Cross-Corpora Language Recognition: A Preliminary Investigation with Indian Languages
Spandan Dey, Goutam Saha, Md Sahidullah

To cite this version:

HAL Id: hal-03223314
https://hal.science/hal-03223314
Submitted on 10 May 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Cross-Corpora Language Recognition: A Preliminary Investigation with Indian Languages

Spandan Dey¹, Goutam Saha¹, Md Sahidullah²
¹Dept. E&ECE, Indian Institute of Technology, Kharagpur, West Bengal, India
²Université de Lorraine, CNRS, Inria, LORIA, F-54000, Nancy, France
sd21@iitkgp.ac.in, gsaha@ece.iitkgp.ac.in, md.sahidullah@inria.fr

Abstract—In this paper, we conduct one of the very first studies for cross-corpora performance evaluation in the spoken language identification (LID) problem. Cross-corpora evaluation was not explored much in LID research, especially for the Indian languages. We have selected three Indian spoken language corpora: IIITH-ILSC, LDC South Asian, and IITKGP-MLILSC. For each of the corpus, LID systems are trained on the state-of-the-art time-delay neural network (TDNN) based architecture with MFCC features. We observe that the LID performance degrades drastically for cross-corpora evaluation. For example, the system trained on the IIITH-ILSC corpus shows an average EER of 11.80 % and 43.34 % when evaluated with the same corpora and LDC South Asian corpora, respectively. Our preliminary analysis shows the significant differences among these corpora in terms of mismatch in the long-term average spectrum (LTAS) and signal-to-noise ratio (SNR). Subsequently, we apply different feature level compensation methods to reduce the cross-corpora acoustic mismatch. Our results indicate that these feature normalization schemes can help to achieve promising LID performance on cross-corpora experiments.

Index Terms—Cross-corpora, language recognition, channel compensation, long-term average spectrum, TDNN.

I. INTRODUCTION

Voice assistants and smart devices are becoming a part of our daily life. Various speech processing applications, such as speech recognition, speech translation, speech synthesis, and speaker recognition, are very important components of these devices [1]. These speech-based applications should have a front-end language identification (LID) module to operate on multiple spoken languages efficiently. This module can predict the spoken language from the speech input and accordingly adapt the mode of operation.

In the last few decades, numerous significant attempts have been made to develop efficient LID systems. Different acoustic, prosodic [2], ASR bottleneck features [3], [4] have been utilized along with several kinds of state-of-the-art classifiers, such as, GMM, i-vector backends [5], and deep neural network based models [6]–[8]. These systems are studied mostly using a single corpus. The test data for evaluating these systems comes from the non-overlapping subsets of the same corpora used for system training. This type of evaluation does not consider the cross-corpora variations, indicating a lack of generalization study for deploying in real-world applications. To the best of our knowledge, cross-corpora study has not been conducted explicitly for spoken language recognition.

Although the cross-corpora study has not been explored explicitly in LID, researchers conducted cross-corpora study in some other speech processing applications, such as anti-spoofing [9] and speech emotion recognition [10]. These works have found that the recognition performance degrades significantly when evaluated with audio-data from other corpora. The main reason for this is the corpora-dependent bias due to differences in data collection methods. Achieving good cross-corpora performance has remained a challenging task.

In Fig. 1, we have shown several factors which can vary considerably across different corpora, collected in different settings for LID task.

In this work, we have conducted one of the very first cross-corpora performance analyses for spoken language recognition with three standard speech corpora in Indian languages. India is a culturally and linguistically diverse country with 1.4 billion population and 22 official languages ¹. For verbal interaction with the smart devices, the major portion of the Indian population is more comfortable with their respective native languages rather than English or other global languages. Due to the mutual influence and similarity among the Indian languages [11], developing Indian LID systems has its unique challenges [12]. Considering these facts, researchers have given special attention to build efficient LID systems [13]–[17] for the Indian languages. However, these systems were developed mostly using a single corpus.

¹https://censusindia.gov.in/2011-common
A. SNR histogram

The SNR histogram helps to analyze the level of background noise present across the speech segments in a speech corpora. We use the NIST STNR tool to calculate the overall signal-to-noise ratio corresponding to each of the speech segments.\footnote{https://www.nist.gov/itl/iad/mig/tools}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{SNR_histogram.png}
\caption{Comparison of SNR histograms for the segments in each corpora.}
\end{figure}

In Fig. 2, histograms of the signal-to-noise ratio values for each of the corpora are shown. Majority of the segments from IIITH corpus have SNR values less than 10 dB, with some segments having SNR values around 20 dB. In contrast, most LDC speech segments have SNR higher than 50 dB, with few low SNR segments. These facts can be justified by the bi-modal nature of the histograms for IIITH and LDC corpora. The histogram for KGP corpus is spread across a wide range of SNR values, indicating more variations in the level of background noise.

B. LTAS analysis

We have compared the corpora in terms of overall spectral information using LTAS analysis. After silence removal, for each segment, the short-time Fourier transform (STFT), $S(f,t) \in \mathbb{R}^{(F \times T)}$ is computed. Here, $F$ denotes the number of frequency bins, and $T$ denotes the time frames. Then, power spectrum $P(f,t)$ is calculated from the STFT output. The LTAS spectrum for a segment $(L_s(f,t))$ is calculated by taking the log of the power spectrum:

$$L_s(f,t) = \frac{1}{T} \sum_{t=1}^{T} \log |S(f,t)|^2$$

Then, overall LTAS for a corpus is computed by taking the average across all its segments. In Fig. 3, the LTAS plots are shown for the training and testing parts of the three corpora. We conclude that the spectral information among the corpora differs significantly at specific frequency ranges, mostly below 100 Hz and above 1 kHz. These ranges of frequency components mainly contribute to the non-speech factors [20], such as channels, background noise, etc.

IV. Methods for Mismatch Reduction

The analysis conducted in Sec. III shows that the mismatch among the corpora is mainly due to varying channel effects and background noise. This section discusses several feature...
compensation schemes to reduce the impact of those variations and apply the same in this work. Among several compensation techniques, we have applied some of the most commonly used ones in the literature, such as CMVN, feature warping, RASTA filtering, along with the relatively newer PCEN approach. Our objective is to prove the effectiveness of these popular techniques for achieving better cross-corpora generalization.

A. Cepstral mean and variance normalization (CMVN)

CMVN assumes that the channel to be static throughout the entire utterance. Thereby, it provides channel-compensated features by making mean zero with mean subtraction followed by variance unity with scaling by the inverted standard deviation. If only mean is subtracted, then this method is called cepstral mean subtraction (CMS). CMVN is expressed as,

$$\hat{X} = \frac{X - \mu}{\sigma}$$

(2)

where, $X \in \mathbb{R}^{N_d \times T}$ is the feature matrix for an utterance with $N_d$ dimensions and $T$ frames, $\mu, \sigma \in \mathbb{R}^{N_d \times 1}$ are the mean and standard deviation respectively across the dimensions. We have also applied windowed CMVN (W-CMVN), where CMVN is applied over windows of three seconds in the utterance. In W-CMVN, the channel is assumed to remain static for a smaller duration, which is a realistic assumption.

B. Feature warping (FW)

Feature warping (FW) warps the short-term distribution of the cepstral features to a standardized distribution [21]. FW increases feature robustness against slowly varying additive noise and various channel and transducer mismatches.

C. Relative spectral (RASTA) processing

RASTA [22] acts as a bandpass filter and removes the static and slowly varying channel effects present in the speech signal. In the high-frequency regions, this bandpass filter also reduces the effect of convolutional noise. Hence, RASTA improves the robustness against environmental variations present in the data. The filter transfer function [22] is expressed as,

$$H(z) = 0.1 * z^4 + \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0.98z^{-1}}$$

(3)

D. Per-channel energy normalization (PCEN)

We have applied PCEN [23] on the mel-spectrograms $(E(f,t))$. During mel-spectral processing, PCEN replaces the logarithmic compression with the following stages [24]: temporal integration (4), adaptive gain control (AGC) (5), and finally dynamic range compression (DRC) (6),

$$M(f, t) = (E * \phi_T)(t, f) = sE(t, f) + (1 - s)M(t - \tau, f)$$

(4)

$$G(t, f) = \frac{E(t, f)}{(M(t, f) + \epsilon)^{\alpha}}$$

(5)

$$PCEN(t, f) = (G(t, f) + \Delta)^{r} - \Delta$$

(6)

Where, $\phi_T(t)$ is a first-order IIR filter with $0 < s < 1$ and $\tau$ as hop size, $0 < \alpha < 1$, $\Delta > 1$, and $r > 0$. PCEN increases the robustness against stationary background noise as well against foreground loudness variations [24].

V. EXPERIMENTAL SETUP

A. Language recognition system

We have computed 20-dimensional mel-frequency cepstral coefficients (MFCCs) features from the silence removed speech using 20 ms frame-size and 10 ms overlap. For classification, we have implemented time-delay neural network (TDNN) based architecture as described in [7] using the PyTorch library [25]. This architecture has five convolutional layers, followed by statistical pooling, and three fully connected layers. We have used categorical cross-entropy as the loss function and AdamW [26] as the optimizer with a learning rate of 0.001. The systems are trained for 30 epochs with validation loss based early stopping criteria and patience of 3 epochs. We follow an end-to-end approach for classification. It gives better validation performance as compared to the backend-based approaches [27]. Using the training data of each of the three corpora, we have trained three independent LID models. Then, each model is evaluated with the test data from all three corpora. If the corpora for training and test are the same, we call this as within-corpora evaluation condition. In contrast, cross-corpora evaluation considers the test data from different corpora than training.

B. Performance evaluation metrics

We have evaluated the LID performance in terms of two standard metrics: equal error rate (EER) [28] and cost average $(C_{avg})$. Using detection error trade-off (DET) plot, the false acceptance and false rejection rates are varied by changing the threshold, and EER is found at the value for which they are equal. In NIST language recognition evaluation (LRE) [29] and OLR challenge [30], $C_{avg}$ was used as the primary evaluation metric. It is defined as follows [29]:

$$C_{avg} = \frac{1}{N} \sum_{L_t} \left\{ \frac{P_{Target} \cdot P_{Miss}(L_t)}{L_t} + \sum_{L_n} P_{Non-Target} \cdot P_{FA}(L_t, L_n) \right\}$$

(7)

where, $L_t$ and $L_n$ are the target and non-target languages. $P_{Miss}$, $P_{FA}$ are the probability of missing and false alarm. $P_{Target} = 0.5$ is the prior-probability of the target languages.
The evaluation results of our baseline experiment are shown in Table II. The results show that the average EER across within-corpora utterances is 11.80%, 27.20%, and 11.18% for the systems trained with IIITH, LDC, and KGP corpus respectively. The relatively higher within-corpora EER for LDC corpus indicates that this corpus contains significant non-lingual variations and is difficult to learn. Cross-corpora evaluations are severely inferior as compared to within-corpora evaluations. The reason for this is the mismatches among the corpora, aroused due to variations in channel and background noise levels, as discussed in Section III. This performance gap shows that LID systems trained on a single corpus can have poor generalization if deployed in real-world scenarios where the environmental variations are even more diverse.

### VI. Results

#### A. Cross-corpora evaluation: baseline experiment

The evaluation results of our baseline experiment are shown in Table II. The results show that the average EER across within-corpora utterances is 11.80%, 27.20%, and 11.18% for the systems trained with IIITH, LDC, and KGP corpus respectively. The relatively higher within-corpora EER for LDC corpus indicates that this corpus contains significant non-lingual variations and is difficult to learn. Cross-corpora evaluations are severely inferior as compared to within-corpora evaluations. The reason for this is the mismatches among the corpora, aroused due to variations in channel and background noise levels, as discussed in Section III. This performance gap shows that LID systems trained on a single corpus can have poor generalization if deployed in real-world scenarios where the environmental variations are even more diverse.

#### B. Effect of feature post-processing

In Table III, we have shown the cross-corpora evaluation for the modified features. The results show that the cross-corpora performances are substantially improved as compared to the baseline performances due to the reduction in mismatch. The within-corpora performances are also improved considerably. For IIITH and KGP trained systems, RASTA processing, and for LDC trained system, PCEN achieve the best cross-corpora improvement. Generally, language recognition performance improves if the test utterance duration is increased [31], [32]. However, for our experiments, this improvement is not very prominent. The TDNN systems are trained with chunks of 3s, and the models are prone to be biased for this 3s duration [33]. In Fig. 4, the performance comparison of the different compensation techniques for the IIITH-trained system and 6s test utterances are shown, using DET plot. The cross-corpora performance gap reduces more effectively between the IIITH and KGP corpus as compared to the LDC corpus. This is because the mismatch is higher between LDC and the other two corpora, which is also justified from Fig. 3. Cross-corpora performance is considerably poor when the KGP-trained systems are evaluated with LDC test set. In these cases, even after the compensation, cross-corpora evaluations show EER more than 50%, which is an interesting fact for further investigation. This indicates that the classifier captures the non-lingual similarities, which affect the similarity score between the target language and test audio [34].
We have studied the cross-corpora performance for spoken language recognition with three corpora. We train three independent LID systems with audio-data from each corpus using a TDNN-based language recognition system. During the evaluation, test data is used from all three corpora. In the baseline experiment, we have shown that the cross-corpora performance is almost around the chance level. We have analyzed that the environmental mismatch is one of the major causes for this. Based on the analysis, we have shown that with the feature level compensation techniques, the corpora mismatch reduces, which leads to a significant improvement in the cross-corpora performance. Among the techniques, CMS and RASTA is found to be more effective for improving generalization. The used compensation techniques can be further improved by tuning several important parameters in a learnable approach. Apart from processing at the feature level, we will also explore the effectiveness of signal level processing. We would also like to solve this cross-corpora problem from domain adaptation perspective in future study.

VII. CONCLUSIONS

Fig. 4. DETs of cross-corpora performances of the compensation techniques when the IIITH corpus is used in training.

REFERENCES