The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages
29-31 August 2018, Gurugram, India

SVM-Based Language Diarization for Code-Switched Bilingual Indian Speech Using Bottleneck Features

Spoorthy V.[1], *Veena Thenkanidiyoor*[1], *Dileep A. D*[2]

[1]National Institute of Technology Goa, India
[2]Indian Institute of Technology Mandi, India

vspoorthy036@gmail.com, veenat@nitgoa.ac.in, addileep@iitmandi.ac.in

Abstract

This paper proposes an SVM-based language diarizer for code-switched bilingual Indian speech. Code-switching corresponds to usage of more than one language within a single utterance. Language diarization involves identifying code-switch points in an utterance and segmenting it into homogeneous language segments. This is very important for Indian context because every Indian is at least bilingual and code-switching is inevitable. For building an effective language diarizer, it is helpful to consider phonotactic features. In this work, we propose to consider bottleneck features for language diarization. Bottleneck features correspond to output of a narrow hidden layer of a multilayer neural network trained to perform phone state classification. The studies conducted using the standard datasets have shown the effectiveness of the proposed approach.

Index Terms: Language diarization, code-switch, bottleneck features, support vector machine.

1. Introduction

India is a multilingual country with more than 600 spoken languages. Every Indian is at least bilingual where he is fluent in his native language and English or Hindi. Code-switching refers to switching of languages in a single conversation. Code-switching is inevitable in Indian speech. It is important to identify the code-switch points in a multilingual speech for the success of automatic speech recognition, spoken dialog systems, machine translation of speech etc. Identifying code-switch points and segmenting a multilingual speech into homogeneous language segments is known as language diarization. Focus of this paper is on language diarization for code-switched bilingual Indian speech involving Kannada and English.

Language diarization and language recognition are two different tasks [1]. This is because, in language recognition, input is a monolingual utterance whereas a multilingual utterance is input in language diarization. An important issue in language diarization is the duration of a language may be very short in a code-switched utterance. A language diarizer inspects a frame of an utterance to decide whether a code-switch occurs in that frame or not. Hence a language diarizer can be built as a binary classifier that discriminates code-switch frames from the rest of the frames. In this work, we propose to build a support vector machine (SVM) based language diarizer.

Early attempts to language diarization explored representing a frame of speech using Mel frequency cepstral coefficients based acoustic features. However soon it was realized that it is essential to use phonotactic information [1]. The approach to language diarization proposed in [1] involved using probabilistic language posterior probabilities. This representation is obtained by building language-specific Gaussian mixture models (GMMs) using the data of respective languages. Speech utterances is represented using frame level posterior probability vectors obtained using language-specific GMMs. To build a language-specific GMM, large amount of data is required. Most of the Indian languages are under resourced. Building a language-specific GMM is a difficult task. To incorporate language-specific features building a language diarizer, we propose to consider bottleneck features in this work.

Bottleneck features are used to parameterize speech signals [11]. Bottleneck features correspond to output of a hidden layer in a neural network where the number of the neurons in the hidden layer is significantly lesser than the surrounding hidden layer. The bottleneck features non-linearly encode the information in the input corresponding to the targets of the neural network. When the targets correspond to phone states, the bottleneck features encode information that is helpful in language diarization. Ideally, such a bottleneck neural network is expected to be trained using the data of both the languages for the bottleneck features to be helpful in developing bilingual language diarizer. The work proposed in this work involves an under resourced language namely, Kannada. We propose to consider a bottleneck neural network that is trained using monolingual English speech [11]. The main idea behind this is when a code-switch utterance is passed through this NN, the network provides relevant information for the English speech whereas it gives irrelevant features for Kannada speech as the network is trained to detect only English phone states. An important contribution of this work is in exploring the use of bottleneck features extracted from a neural network trained using monolingual English speech for developing language diarizer for bilingual Indian speech. Availability of datasets for Indian languages is difficult. Another important contribution of this paper is in building a code-switch dataset involving Kannada and English. The studies conducted in this work show the effectiveness of the proposed approach.

This paper is organized as follows: In Section 2, we present related work. The proposed work is presented in Section 3. The experimental studies are presented in Section 4. In Section 5, conclusions are presented.

2. Related work

An approach to language diarization for the Mandarin and English languages is proposed in [1]. This includes acoustic and phonotactic features for language diarization and segmentation. In this work, it was observed that code-switching can occur very often and the average interval of a language can be of a short span. For processing these short segments,the proposed language diarization system included long term context feature covering many phone-based segments. In [2], a language diarization approach for Frisian and Dutch languages is proposed. In this approach, impact of a bilingual Deep Neural

10.21437/SLTU.2018-28

Network (DNN) in the context of code-switching speech, an automatic speech recognition is investigated. The detection of language boundaries for code-switch speech utterances for languages such as Swahili and English have been proposed in [6]. In [7] an approach to identify the language boundary in code-switching utterances with the use of bi-phone probabilities is explored. A comparative study on selecting acoustic modeling units in deep neural networks based large vocabulary Chinese speech recognition was proposed in [8]. As per our knowledge, there is no attempt in language diarization in the context of Indian languages. Our focus is to build a language diarization system that is able to identify the code-switch points in a bilingual code-switch conversational speech in Indian languages. In this work, code-switched speech in English and Kannada languages is considered.

3. Proposed Approach

The task of language diarization involves correctly detecting code-switch points when more than one language is used by a speaker. In Figure 1, a bilingual code-switch speech utterance is plotted where portions uttered in a particular language is plotted in single color. There exists a high degree of similarity among the frames of speech involving a code-switch point such as X shown in Figure 1 and the frames of speech not having the code-switch frames. Hence, we need an effective discriminative classifier that discriminates code-switch frames from non-code-switch frames. A language diarizer can be built as a binary classifier that discriminates a code-switch frame from non code-switch frames. In this work, we propose to consider support vector machines(SVMs) that are found to be effective discriminative classifier to identify code-switch points. Since a code-switch point (for example X shown in Figure 1) is part of multiple frames of speech considered for feature extraction, we propose to consider L frames around the point X as the positive examples for code-switch frames.

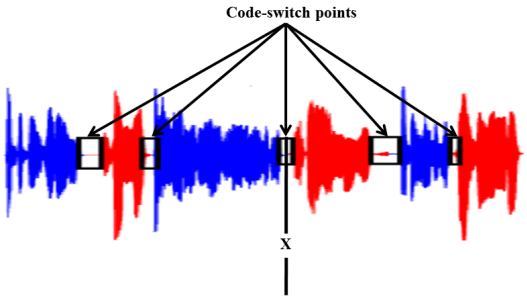

Figure 1: *An illustration of code-switch utterance and the corresponding code-switch points.*

To build an effective SVM-based language diarizer, it is necessary is represent the frames in a suitable manner. In this work, we propose to consider the bottleneck features to represent a frame. The bottleneck features correspond to the recent development in speech signal parameterization. Bottleneck features are extracted from a hidden layer of a multilayer feed forward neural network where the hidden layer will have significantly lesser number of neurons when compared to the surrounding hidden layers. Such a bottleneck neural network is illustrated in Figure 2. In the multilayer neural network shown in Figure 2, there are 5 hidden layers where the 3^{rd} one is having significantly lesser number of neurons. This layer is known as the bottleneck layer [11]. The neural networks of the type shown in Figure 2 are trained to map the input to output. The bottleneck features representation is taken as a by-product of neural network trained to do such mapping. The bottleneck features encode information in speech signal corresponding to K targets considered. When the targets correspond to phone states the bottleneck features nonlinearly encode the phonetic information [11]. In this work, we propose to consider a bottleneck neural network whose targets are phone states.

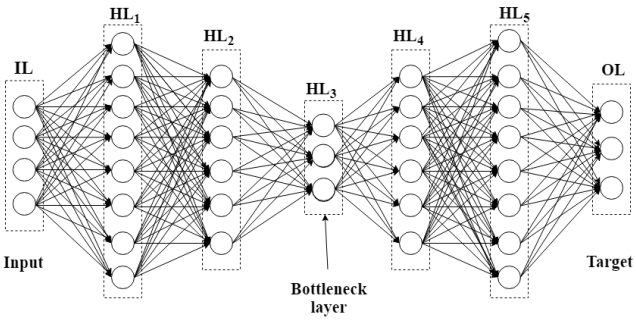

Figure 2: *Illustration of bottleneck neural network where the third hidden layer has significantly lesser number of neurons than the other hidden layers. The output layer comprises of K neurons corresponding to the targets. Here, IL stands for input layer, HLi stands for ith hidden layer and OL corresponds to output layer.*

4. Experimental Studies

In this section, we present the experimental studies conducted. We first present the datasets considered for the studies. Then we present the representations considered for speech utterances following which we present the experimental studies.

4.1. Datasets

In this work, we consider bilingual code-switch speech involving an Indian language namely Kannada and English. In this work, we consider the following two datasets for the studies.

NIT Goa Bilingual Code-Switch Corpora: In this corpora, 7 speakers from NIT Goa have been asked to utter a set a code-switch sentences. Each speaker had been instructed to read out 10 code-switch sentences. The speakers chosen for the recording are native speakers of Kannada. The corpora consists of 70 code-switch utterances. For the evaluation, 10-fold stratified sampling have been performed and the accuracies presented correspond to average of the accuracy obtained for 10 folds. To train the system, 7 randomly chosen code-switch utterances from each speaker is considered and the rest is used for testing. Due to lack of native speakers at the Institute, this corpora is of limited size.

Artificially Generated Code-Switch Corpora: In this corpora, the speech samples of English and Kannada are taken from the dataset proposed in [12]. The utterances in this dataset were recorded in a noise-free studio by two speakers, 1 male and 1 female. The dataset consisted of 15 hours of English data and 18.04 hours of Kannada data. In this work, 70 artificially stitched code-switch utterances are considered. For code-switched utterances, random audio files are selected from both English and Kannada utterances and artificially stitched to cre-

ate code-switched utterances. For the evaluation, 10-fold stratified sampling have been performed and the accuracies presented correspond to average of the accuracy obtained for 10 folds.
In both the datasets, code-switch points in an utterance is marked manually at a specific millisecond and then mapped to that respective frames. Let f denote the code-switch frame and in our experiment we considered f+5 and f-5 frames as the code-switch frames. Rest of the frames, irrespective of their languages are considered as non code-switched frames.

4.2. Representations

Every utterance is sampled at 8000 Hz before extracting features from them. In this work, we represent every utterance using Mel frequency cepstral coefficients(MFCC), language posterior probabilities and bottleneck features. MFCC features are most commonly used acoustic features. These features consider human perception sensitivity taking frequencies into consideration. For extraction of MFCC features from the code-switch utterances, a frame size of 25 milliseconds and frame shift of 10 milliseconds is considered. A total number of 40 filters and 13 cepstral coefficients are extracted from each frame. To represent a frame of speech using language posterior probabilities, Gaussian mixture models (GMMs) for Kannada and English are built. The number of components for the GMM is chosen by systematically varying it from 32 upto 1024 and the suitable one is chosen. Every frame of speech is represented using the two dimensional vector of posterior probabilities obtained from the language specific GMMs.

In this work, to extract bottleneck features from an audio signal, Phonexia bottleneck feature extractor tool [11] is used. This network comprises of a bottleneck layer of 80 neurons. The training corpora used for this network is Fisher English [12] with 120 phoneme states as output classes (40 phonemes, 3 state for each phoneme). The neural network is trained for monolingual phone state classification. Though it is expected to train the bottleneck neural network using data of Kannada and English for the bottleneck features to be effective in developing a language diarizer for bilingual code-switch speech, we propose to use a neural network trained using monolingual data. This is mainly because of the non availability of sufficient data for an under resourced language like Kannada. The main idea behind this is when a code-switch utterance is passed through this neural network, the network provides relevant information for the English speech whereas it gives irrelevant features for Kannada speech as the network is trained to detect only English phone states. This is expected to be helpful in discriminating code-switch frames from non code-switch frames. To extract bottleneck features from the code-switch utterances, a sampling frequency 8000 Hz, a frame size 25 milliseconds, a frame shift of 15 milliseconds, high frequeny of 3800 Hz, low frequeny 64 Hz, and 24 channels are considered.

To consider the temporal information in speech, we consider the context information in this work. For MFCC features use of contexts 1,2,3,4 and 5 are explored and the suitable one is chosen. In case of language posterior probability representation, for the context of 1, posterior probability vectors corresponding to 3 MFCC feature vectors are concatenated to get a 6 dimensional vector. Similarly, for the MFCC features of context N, posterior probability vector of 2((N*2)+1) dimension is obtained. For bottleneck features, a global context of 15 is considered in the bottleneck neural network.

4.3. Studies on Artificially Generated Bilingual Code-Switch Speech Corpora

In this work, we propose to build an SVM-based language diarizer. The SVM-based language diarizer is trained to discriminate the code-switch frames from non code-switch frames. For SVM-based language diarizer in this work, we have considered two kernels namely linear kernel and Gaussian kernel. The trade off parameter C and kernel width parameter of Gaussian kernel are chosen emperically. SVM-based language diarizer is built for the Artificially Generated Code-Switch Speech Corpora by using the MFCC features, language posterior probability features and bottleneck features respectively. The context of the frames for the MFCC feature vectors are varied from 1 to 5 and the best accuracy was obtained for a context of 4. In this work, we consider a context of 4 for MFCC feature vectors for the rest of the studies. The accuracies for SVM-based language diarization using MFCC features is given in Table 1. MFCC features correspond to acoustic information of a speech signal. To perform language diarization, language-specific features may be helpful. To consider language specific features, language posterior probability feature vector representation of a frame of speech is considered. In this work, we propose to build a language-specific GMM each for Kannada and English using the respective speech data from the dataset proposed in [12]. MFCC feature vectors of the code-switch utterances are passed to each GMM. The posterior probability generated from both English and Kannada GMMs is taken as a 2 dimensional posterior probability vector. A context of 4 is considered for the language posterior probability representation also. The accuracies for the SVM-based language diarization using the posterior probability vector representation using a context of 4 is given in Table 1. In this work, we also propose to consider combined MFCC and language posterior probability features to represent a frame of speech. For this, the MFCC feature vector with context is concatenated with the language posterior probability vector. The accuracies for the combined representation is given in Table 1. To bring more language specific information, we propose to extract bottleneck features for the code-switch utterances. These features provide phonotactic information for the code-switch utterances. The accuracies obtained for the SVM-based language diarization using the bottleneck features is given in Table 1. The performance of SVM-based language diarizer using difefrent representations is also compared in Figure 3. It is seen from Table 1 and Figure 3 that the SVM-based language diarizer using Gaussian kernel performs better than that uses Linear kernel. It is also seen from Table 1 and Figure 3 that accuracy of SVM-based language diarizer using bottleneck featutes is better than that obtained using other representations. This shows the effectiveness of using the bottleneck features. Next, we present our studies on NIT Goa Bilingual Code-Switch Speech Corpora.

4.4. Studies on NIT Goa Bilingual Code-Switch Speech Corpora

For this dataset also, we propose to build an SVM-based language diarizer. In this work, we have considered two kernels namely linear kernel and Gaussian kernel. The trade off parameter C and the kernel width parameter of Gaussian kernel are chosen emperically. SVM-based language diarizer is built for the NIT Goa Code-Switch Speech Corpora by using the MFCC features, language posterior probability features and bottleneck features respectively. For MFCC feature vector representation, we have considered a context of 4. The accuracies of lan-

Table 1: *Accuracy for SVM-based language diarizer using MFCC, language posterior probability and bottleneck feature representations for speech. Here we consider a context of 4 for MFCC and language posterior probability vectors and a global context of 15 for bottleneck features. Accuracies presented here correspond to average of accuracies obtained for 10 folds.*

Representation	Linear kernel	Gaussian kernel
MFCC	70.72%	73.03%
Language PP	58.24%	61.05%
MFCC+Language PP	71.08%	69.18%
Bottleneck features	83.96%	86.16%

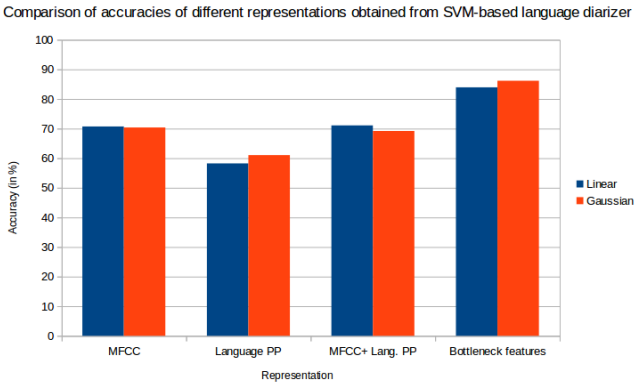

Figure 3: *Comparison of different representations of code-switch utterance of Artificially Generated bilingual Code-Switch Speech Corpora.*

guage diarizer using MFCC features is given in Table 2. MFCC features provide only acoustic information for a speech signal. To perform language diarization, language-specific features are helpful. To consider language specific features, posterior probability feature vector representation of a frame of speech is considered. Since the NIT Goa bilingual code-switch speech corpora is of limited size, it is difficult to build language specific GMM. Hence in this work, we propose to adapt the language-specific GMM built using the Kannada and English data of TTS dataset [12] as presented in Section 4.3 to the English and Kannada data of NIT Goa Bilingual Code-Switch Speech Corpora respectively. MFCC feature vectors of the code-switch utterances are passed to each adapted GMMs. The posterior probability generated from both the adapted GMMs is taken as a 2 dimensional posterior probability vector and a context of 4 is considered for language posterior probability representation. The accuracies for the SVM-based language diarization using the posterior probability vector representation using a context of 4 is given in Table 2. In Table 2, we also give the performance of SVM-based language diarizer using combined MFCC and language posterior probability representation. As mentioned earlier, to bring more language specific feature, we propose to extract bottleneck features for the code-switch utterances. The accuracies obtained for the SVM-based language diarization using the bottleneck features is given in Table 2. The performance of SVM-based language diarizer using difefrent representations is also compared in Figure 4. It is seen from Table 2 and Figure 4 that the accuracy of SVM-based language diarizer using bottleneck features is better than that obtained using MFCC features and the language posterior probability features. This shows the

Table 2: *Accuracy for SVM-based language diarizer using MFCC, language posterior probability and bottleneck feature representations for speech. Here we consider a context of 4 for MFCC and language posterior probability vectors and a global context of 15 for bottleneck features. Accuracies presented here correspond to average of accuracies obtained for 10 folds.*

Representation	Linear kernel	Gaussian kernel
MFCC	81.26%	78.37%
Language PP	69.88%	65.08%
MFCC+Language PP	59.79%	61.59%
Bottleneck features	82.86%	83.16%

effectiveness of the proposed approach.

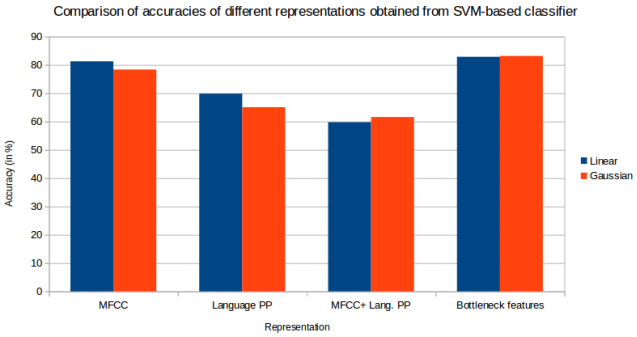

Figure 4: *Comparison of different representations of code-switch utterance of NIT Goa bilingual Code-Switch Speech Corpora.*

5. Conclusions

In this work, we proposed a SVM-based language diarizer for bilingual code-switch speech involving an under resourced language, Kannada and English. For building the language diarizer, we proposed the use of bottleneck features . The proposed language diarizer using bottleneck features was found to be better than the language diarizer built using other features. This shows the effectiveness of the use of bottleneck features. In the present work, a bottleneck network trained on monolingual data is used. In future, we propose to explore the possibility of training the bottleneck neural network using bilingual data. In future, the proposed approach can also be extended to other Indian languages. This is very helpful because India is a multilingual country and code-switching is inevitable.

6. References

[1] Lyu DC, Chng ES, Li H., "Language diarization for code-switch conversational speech",IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013 (2013 May 26),pp–7314-7318.

[2] Ylmaz E, van den Heuvel H, van Leeuwen D.: Investigating bilingual deep neural networks for automatic recognition of code-switching frisian speech. Procedia Computer Science. (2016 Jan 1);81:159-66.

[3] Anguera X, Bozonnet S, Evans N, Fredouille C, Friedland G, Vinyals O.: Speaker diarization: A review of recent research. IEEE Transactions on Audio, Speech, and Language Processing. (2012) Feb;20(2):356-70.

[4] Besacier L, Barnard E, Karpov A, Schultz T.: Automatic speech recognition for under-resourced languages: A survey. Speech Communication. (2014), vol. 56, pp–85-100.

[5] Gray S, Hansen JH. : An integrated approach to the detection and classification of accents/dialects for a spoken document retrieval system. In Automatic Speech Recognition and Understanding, 2005 IEEE Workshop on (2005 Nov 27), pp– 35-40. IEEE.

[6] Kleynhans N, Hartman W, Van Niekerk D, Van Heerden C, Schwartz R, Tsakalidis S, Davel M. : Code-switched English pronunciation modeling for Swahili spoken term detection. Procedia Computer Science. (2016 Jan 1). 81:128-35.

[7] Chan JY, Ching PC, Lee T, Meng HM. : Detection of language boundary in code-switching utterances by bi-phone probabilities. In Chinese Spoken Language Processing, 2004 International Symposium on (2004 Dec 15), pp–293-296.

[8] Li X, Yang Y, Pang Z, Wu X. : A comparative study on selecting acoustic modeling units in deep neural networks based large vocabulary Chinese speech recognition. Neurocomputing (2015), Dec 25; vol. 170:251-6.

[9] Li H, Ma B, Lee KA. : Spoken language recognition: from fundamentals to practice. Proceedings of the IEEE. (2013), vol. 101(5), pp–1136-59.

[10] Lafferty J, McCallum A, Pereira FC. : Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

[11] Fr R, Matjka P, Grzl F, Plchot O, Vesel K, ernock JH. : Multilingually trained bottleneck features in spoken language recognition. Computer Speech and Language, vol. 46, pp–252-67 (Nov 1, 2017). .

[12] Baby A., A. L. Thomas and H. A. Myrthy. : Resources for Indian languages. Proceedings of Text, Speech and Dialogue. (2016) .