Received July 25, 2020, accepted August 20, 2020, date of publication August 31, 2020, date of current version October 5, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3020506

Mel-Weighted Single Frequency Filtering Spectrogram for Dialect Identification

RASHMI KETHIREDDY[1], (Graduate Student Member, IEEE), SUDARSANA REDDY KADIRI[2], (Member, IEEE), PAAVO ALKU[2], (Fellow, IEEE), AND SURYAKANTH V. GANGASHETTY[1], (Member, IEEE)

[1]Speech Processing Laboratory, International Institute of Information Technology-Hyderabad (IIIT-H), Hyderabad 500032, India
[2]Department of Signal Processing and Acoustics, Aalto University, 00076 Espoo, Finland

Corresponding author: Rashmi Kethireddy (rashmi.kethireddy@research.iiit.ac.in)

The work of Rashmi Kethireddy was supported by the University Grants Commission, India, under Project 3582/(NET-NOV2017). The work of Sudarsana Reddy Kadiri was supported by the Academy of Finland, Projects, under Grant 312490 and Grant 330139.

ABSTRACT In this study, we propose Mel-weighted single frequency filtering (SFF) spectrograms for dialect identification. The spectrum derived using SFF has high spectral resolution for harmonics and resonances while simultaneously maintaining good time-resolution of some speech excitation features such as impulse-like events. The SFF spectrum can represent speech characteristics such as burst time and glottal closure instances better than the short-time Fourier transform (STFT) spectrum. Our hypothesis is that these intricate representations in the SFF spectrum should help in distinguishing dialects. Therefore, we built a dialect identification system which uses an unsupervised, bottleneck feature representation of the Mel-weighted SFF spectrogram (Mel-SFF spectrogram) with sequence-to-sequence deep autoencoders. The language invariance of the proposed system was evaluated using two datasets: the UT-Podcast database (English) and the STYRIALECT database (German). The proposed representations gave a relative improvement of 9.47% and 4.69% in unweighted average recall (UAR) compared to the best baseline method on the development and test datasets, respectively, of the UT-Podcast database. The proposed representations also gave a comparable performance to the best baseline method for the STYRIALECT database. In addition, the fusion of the autoencoder bottleneck features computed from the Mel-SFF and Mel-STFT spectrograms improved the overall performance indicating complementary information between these features. By further analyzing the performance of the proposed representation with different utterance lengths using the UT-Podcast database, we observed that the proposed representation performed better on short utterances. The improved performance given by the Mel-weighted SFF spectrogram for recognizing dialects in both databases supports our hypothesis.

INDEX TERMS Dialect identification, single frequency filtering (SFF) spectrum, Mel-spectrogram, Mel-filter bank energies, autoencoder.

I. INTRODUCTION

In listening to speech, humans not only analyse the speech signal's linguistic content but they also make conclusions about the speaker's regional origin, social background and emotional state. Dialect identification refers to a research area where the goal is to find the regional origin of the speaker using the temporal and spectral characteristics of his or her speech signal. Each dialect group has its own pronunciation pattern and vocabulary compared to other dialect groups. These variations in speech due to dialect have been shown to decrease the performance of automatic speech recognition (ASR) systems. An efficient dialect identification system followed by a dialect-specific pronunciation dictionary and a dialect-specific language model can improve the performance of ASR [1]–[3]. In addition, dialect information can be used in speaker profiling in biometrical applications, it can help solve dialect related issues in speaker and language identification, and it can also be used in the development of dialect-personalized voice assistants.

Dialect identification studies in the literature can be classified into two groups: text-dependent and text-independent [4]. In the former, the transcription of an utterance for which the dialect needs to be identified should be known apriori.

The associate editor coordinating the review of this manuscript and approving it for publication was Jing Liang.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

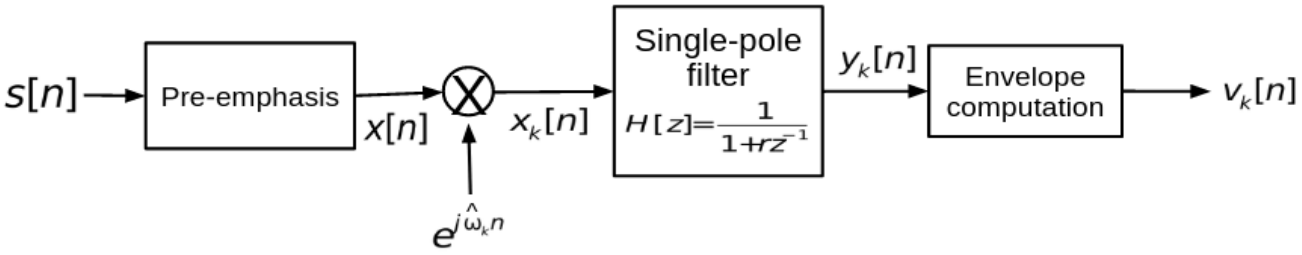


FIGURE 1. Block diagram describing the steps involved in the single frequency filtering (SFF) method [33].

The dialect is determined by finding the closest dialect match for a word/utterance by either phone modeling or word modeling. Phone/word sequences are modeled using n-gram models. For the dialect identification task, different phone modeling approaches are widely used and they include phone recognition followed by language modeling (PRLM) and parallel phone recognition followed by language modeling (PPRLM) [5]–[8].

Dialect identification studies belonging to the second class, text-independent dialect identification, model dialectal variations using acoustic features derived from speech signals [9]–[13]. The acoustic features used in this area include shifted delta cepstral coefficients (SDCs) [14], prosody based features, frame-by-frame phone posteriors, supervectors [15]–[17], and i-vectors obtained from acoustic features and speech attributes [18]–[21]. In one of the recent studies in language identification, bottleneck features (BNFs) derived from a pre-trained deep neural network with i-vector modeling showed significant improvement over the SDC features, and the developed BNF-based system stands out as state of the art [22], [23]. A shortcoming of this approach is that the deep neural network had to be trained over a transcribed corpus which contains only English speech without phonetic variations in pronunciation [24].

This article studies text-independent dialect identification without taking advantage of any pre-trained models or transcriptions. The proposed system takes advantage of an autoencoder which is trained using the Mel-weighted single frequency filtering (Mel-SFF) spectrogram to obtain BNFs which are used in the classification. A baseline system with a similar architecture is trained using the Mel-weighted short-time Fourier transform (Mel-STFT) spectrogram. The autoencoder architecture used is similar to the one developed in [25], [26]. This autoencoder model converts a variable-length feature vector to a fixed-length representation in an unsupervised manner. This architecture was chosen in the current study because it was shown in [25] to be the best performing system in dialect classification compared to two reference techniques. The spectrum computed by single frequency filtering (SFF) has been shown to give good spectral resolution to indicate harmonics and resonances [27] and good temporal resolution to model speech excitation features such as impulse-like events [28]. The SFF spectrum has also shown promising performance in determining burst-onset points related to voice-onset time (VOT) and glottal closure instances compared to the short-time Fourier transform (STFT) spectrum [28]–[30]. Previous studies in dialect identification have shown the significance of VOT for identification of accent [31]. Inspired by this, we propose to use the Mel-weighted SFF spectrogram with autoencoders to derive fixed-length speech representations for dialect identification.

The organization of the paper is as follows: Section II describes the SFF method and the computation of the Mel-weighted SFF spectrogram. Section III provides a detailed description of the proposed dialect identification system. The experimental setup is described in Section IV. Results are presented in Section V. Finally, Section VI summarizes the study.

II. SFF AND COMPUTATION OF THE MEL-WEIGHTED SFF SPECTROGRAM

This section describes the steps involved in the SFF method and in the computation of the Mel-weighted SFF spectrogram.

A. SFF

The SFF method is used to derive the amplitude envelope of the speech signal at every sample for a given frequency [32]. The SFF spectrum has been shown to be useful in finding burst-onset points [29] and glottal closure instants [30], and it has been demonstrated to exhibit high spectral resolution for important speech features such as harmonics and resonances [27].

In SFF, the pre-emphasized speech signal is used for deriving the amplitude envelope at each frequency by frequency-shifting the signal and by filtering it using a single-pole filter as shown in Figure 1. The pole of the filter is located on the negative real axis close to the unit circle in the z-plane, i.e., the angle of the pole corresponds to the Nyquist frequency ($\frac{fs}{2}$). Therefore, the effect of other frequency components will be reduced giving high spectral resolution. The steps to derive the SFF spectrum are given below [32].

- Speech signal (s[n]) is pre-emphasized to remove low-frequency variations. The pre-emphasis is computed as follows

$$x[n] = s[n] - \alpha * s[n-1], \tag{1}$$

where α is set to 0.95 in the present study.

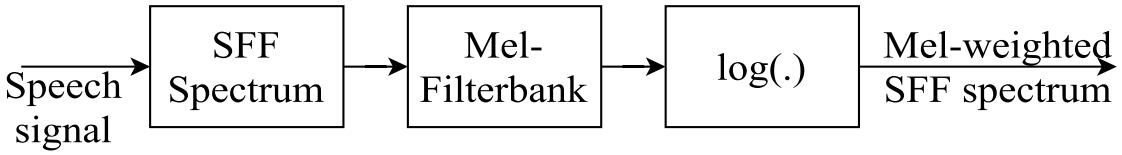


FIGURE 2. Block diagram describing the steps involved in the computation of the Mel-weighted SFF spectrum. The spectrogram obtained using this process refers to the Mel-weighted SFF spectrogram.

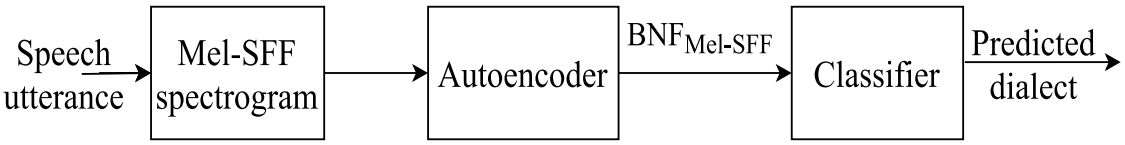


FIGURE 3. Block diagram of the proposed dialect identification system.

- The pre-emphasized speech signal $x[n]$ is multiplied by complex sinusoid $e^{j\hat{\omega}_k n}$ as follows

$$x_k[n] = x[n]e^{j\hat{\omega}_k n}, \tag{2}$$

where $\hat{\omega}_k = \pi - \frac{2\pi f_k}{f_s}$, f_k is the desired frequency and f_s is the sampling frequency.
- Signal $x_k[n]$ is passed through a single-pole filter. The transfer function of the filter is defined as

$$H(z) = \frac{1}{1 + rz^{-1}}, \tag{3}$$

where $r \approx 1$, i.e., the pole is close to the unit circle on the negative real axis in the z-plane. In this study, r is set to 0.99.
- The output of the filter is given by

$$y_k[n] = -ry_k[n-1] + x_k[n]. \tag{4}$$

- The amplitude envelope ($v_k[n]$) of the signal with the desired frequency f_k is given by

$$v_k[n] = \sqrt{y_{kr}^2[n] + y_{ki}^2[n]}, \tag{5}$$

where $y_{kr}[n]$ and $y_{ki}[n]$ are the real and imaginary parts of $y_k[n]$.

The amplitude envelope can be computed for several frequencies using a frequency interval (Δf) as follows

$$f_k = k\Delta f, \quad k = 1, 2, \ldots K \tag{6}$$

where $K = \frac{(\frac{f_s}{2})}{\Delta f}$. In this study $K = 2048$ is used, i.e., $\Delta f = \frac{(\frac{f_s}{2})}{2048}$. From the amplitude envelope $v_k[n]$, the SFF spectrum of the signal is obtained at each instant of time.

B. COMPUTATION OF THE MEL-WEIGHTED SFF SPECTROGRAM

This section describes the computation of the Mel-weighted SFF spectrogram. The procedure, depicted in Figure 2, consists of the extraction of the filter-bank energies obtained by filtering the SFF spectrum with triangular Mel-spaced filters followed by logarithm. The resulting Mel-filter bank energies (MFBE) are referred to as MFBE-SFF or simply as the Mel-weighted SFF spectrum. For convenience, we refer to the spectrogram obtained by using this process as the Mel-weighted SFF spectrogram or Mel-SFF spectrogram.

As explained in Section II-A, SFF provides the spectrum at each instant of time. Instead of considering the spectrum at each time instant, computational load is reduced in the current study by considering the spectrum unchanged in a segment of T ms. One of the following four approaches can be used in defining the spectrum using the segment of T ms.

(a) **Average SFF spectrum ($\mathbf{S_{avg}}$)**: In this approach, the SFF spectrum is computed by averaging the amplitude envelope $v_k[n]$ defined in Eq. 5 for every frequency k over the entire segment.
(b) **Minimum SFF spectrum ($\mathbf{S_{min}}$)**: In this approach, the SFF spectrum is selected as the instantaneous spectrum of $v_k[n]$ which shows the minimum spectral energy (sum of the squared amplitude envelope values) over the entire segment.
(c) **Maximum SFF spectrum ($\mathbf{S_{max}}$)**: In this approach, the SFF spectrum is selected as the instantaneous spectrum of $v_k[n]$ which shows the maximum spectral energy over the entire segment.
(d) **Uniform SFF spectrum ($\mathbf{S_{uniform}}$)**: In this approach, the SFF spectrum is computed by sampling $v_k[n]$ at regular intervals defined by the segment duration.

The performance of the above four approaches was compared in this study. As will be reported in Section V, it was observed that S_{avg} gave the best performance. Therefore, the Mel-SFF spectrogram computed using S_{avg} was used as a spectral representation of speech and this representation was further processed in an unsupervised manner by an autoencoder to obtain fixed-sized BNFs to be used in dialect identification.

III. PROPOSED SYSTEM

The proposed system has three stages: the Mel-SFF spectrogram extraction, obtaining an unsupervised representation from spectrograms using an autoencoder and classification. The block diagram shown in Figure 3 describes the proposed system architecture. For convenience, we refer to the unsupervised representation from the STFT spectrogram as

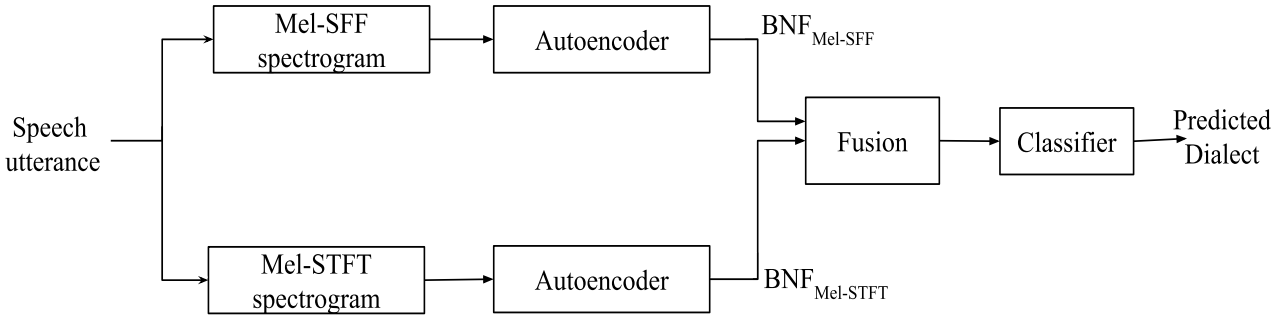


FIGURE 4. Block diagram of the proposed fusion dialect identification system.

$BNF_{Mel\text{-}STFT}$ and the unsupervised representation from the SFF spectrogram as $BNF_{Mel\text{-}SFF}$.

In addition, we also studied a system using feature fusion where the unsupervised representations (BNFs) derived from Mel-SFF spectrograms ($BNF_{Mel\text{-}SFF}$) were combined with the unsupervised representations (BNFs) derived from conventional Mel-STFT spectrograms ($BNF_{Mel\text{-}STFT}$). The fusion system, depicted in Figure 4, is discussed in Section III-C.

A. MEL-SFF SPECTROGRAM CONFIGURATION

The time-domain speech signal is processed as described by Eqs. 1-5 in Section II to compute the SFF spectrogram. Instead of considering the spectrum at every sample, averaging of the spectrogram is computed as explained in Section II-B. We varied the average energy operation (S_{avg} explained in Section II-B) by considering seven values of T (6.25 ms, 12.5 ms, 18.75 ms, 25 ms, 31.25 ms, 37.5 ms, 43.75 ms). The best performance was obtained with $T = 37.5$ ms. Therefore, this segment duration value is used throughout this study unless otherwise mentioned. Mel-filter bank energies are obtained from the spectrum using linearly spaced 128 filters in the Mel-scale. To eliminate the effect of background noise, the spectral values are clipped using a threshold as in [25]. In this study, five thresholds (−30 dB, −40 dB, −50 dB, −60 dB, and −70 dB) are explored. Spectrograms are normalized to be in the range of [−1, 1] to match with the decoder output function in the autoencoder. The clipping and normalization operations are conducted in the same way in computing the proposed Mel-SFF spectrogram and in computing its reference spectrogram, the conventional Mel-STFT spectrogram.

B. SEQUENCE-TO-SEQUENCE AUTOENCODERS

Sequence-to-sequence autoencoders compress high-dimensional frame-level representations to low-dimensional utterance-level latent representations by capturing the relevant information to reproduce the input sequence. The low-dimensional utterance-level representations capture the required information compactly, which can be used for classification. A sequence-to-sequence recurrent neural network (RNN) with an autoencoder framework is used to convert the Mel-SFF spectrogram to an utterance-level fixed representation. The autoencoder framework has two modules, the encoder and the decoder. In both modules, sequence-to-sequence RNNs are used. Motivated by [26], we used gated recurrent cell units (GRU) throughout the study.

The encoder converts the Mel-SFF spectrogram to a fixed-length representation. A fully-connected layer with a *tanh* activation function converts the output of the encoder to a hidden input format which is considered the BNF. The BNFs extracted from the trained autoencoder are used for dialect identification.

The BNFs are passed to the decoder as hidden inputs and the decoder learns to reproduce the Mel-SFF spectrogram sample by sample. The estimated output is recurrently passed to the next time-steps as hidden state input. The autoencoder network is trained by minimizing the root mean square error (RMSE) between the estimated decoder output and the original Mel-SFF spectrogram. The initial hidden state input for the first time steps of the RNNs in the encoder and the initial input in the decoder are set to 0 for all utterances. The autoencoders of this study are implemented using the AuDeep toolkit [26], [34].

C. FUSION SYSTEM

In order to investigate whether there is complementary information between the STFT and SFF spectrograms, we developed a fusion system. The block diagram of the fusion system is shown in Figure 4. In this system, the bottleneck features extracted from the autoencoders trained on the Mel-SFF spectrogram (the proposed $BNF_{Mel\text{-}SFF}$ system) and Mel-STFT spectrogram (the baseline $BNF_{Mel\text{-}STFT}$ system) are concatenated to train the classifier. Two separate autoencoders are trained, each one capturing the underlying latent space representations from respective input spectrograms. The classifier trained using these fused features is used for dialect identification in a similar manner as the system described in Section III-D.

D. CLASSIFICATION

The third stage in the proposed system is classification. We experimented with three different classifiers:

Gaussian linear classifier (GLC), multi-class logistic regression (MCLR) and support vector machine (SVM). GLC is a generative classifier model and MCLR as well as SVM are discriminative classifier models. GLC was implemented based on [35]. For MCLR and SVM, we used the implementations from [36]. Both MCLR and SVM use the one-vs-rest strategy to classify dialects.

IV. EXPERIMENTAL SETUP

In this section, the databases used in the study are described. In addition, the section discusses the baseline systems and the evaluation metrics adopted in the study.

A. DATABASES

Two speech databases, STYRIALECT and UT-Podcast, are used in the study. The STYRIALECT database includes the dialects of Styria in German [25]. The database consists of 5227 utterances for training and 2570 utterances for development. The STYRIALECT test set is not provided with labels and therefore the results of this database are reported only for the development set in this study. The sampling frequency is 16 kHz and the duration of each utterance is 2 *s*. The database has three classes (Styrian dialects of German) with different distributions. More details about the database can be found in [25].

The UT-Podcast database consists of three major dialects (US, UK, and AU) of English [37]. This data is collected from different websites for each dialect and it covers a wide range of topics. Since the data is collected from online podcasts, speech is more spontaneous than in STYRIALECT and not very well structured. Therefore, the collected speech captures all the dialectal traits (pronunciation, vocabulary, and grammatical variations). The speech signals are segmented in such a way that each utterance is 17 *s* in duration and contains 46 words on average. The sampling frequency is 8 kHz. The database is divided into train and test sets as described in [37]. For the experiments in this study, half of the original test set of the database is used for development and the other half for testing.

B. BASELINE CONFIGURATIONS

The proposed system is compared to three baseline systems. The first baseline is the ComParE'19 system, which uses the BNFs of a sequence-to-sequence autoencoder, which is trained using the Mel-STFT spectrogram [25]. This system will be referred to as the $\text{BNF}_{\text{Mel-STFT}}$ system. The second baseline is an i-vector system which is trained using Mel frequency cepstral coefficients (MFCCs) [20], [38]. The third baseline is an i-vector system, which is trained using the BUT/phonexia bottleneck features [39]. The second and third system will be referred to as the $\text{i-vector}_{\text{MFCC}}$ system and the $\text{i-vector}_{\text{BUT-BNF}}$ system, respectively.

In the $\text{BNF}_{\text{Mel-STFT}}$ system, spectral analysis of speech is computed with STFT using 80-*ms* frames with the Hann window and a shift of 40 *ms*. Mel-bank energies are computed from the spectrum using 128 channels. Amplitude clipping is performed on the Mel-STFT spectrogram to reduce the effect of noise captured in the recordings. Five clipping thresholds, denoted as −40 dB, −50 dB, −60 dB, −70 dB, and −80 dB, are generated as in [25]. The baseline $\text{BNF}_{\text{Mel-STFT}}$ system uses a similar autoencoder architecture as the proposed $\text{BNF}_{\text{Mel-SFF}}$ system in order to make a fair comparison between the use of the two Mel-weighted spectrograms in dialet identification. The RNNs in the autoencoder have two layers with 256 GRUs in each layer and the encoder network is unidirectional while the decoder network is bidirectional. The network is trained for 16 epochs with a drop out of 30%. The $\text{BNF}_{\text{Mel-STFT}}$ system is trained in a similar manner as in the proposed $\text{BNF}_{\text{Mel-SFF}}$ system to obtain unsupervised representations (BNFs) from the Mel-weighted spectrograms. These representations (BNFs) are used to train the classifier, which is then used for the dialect prediction.

The other two baseline systems (the $\text{i-vector}_{\text{MFCC}}$ system and the $\text{i-vector}_{\text{BUT-BNF}}$ system) differ only in the feature representations used for i-vector training. In the former, i-vectors are extracted from 13 static mean normalized MFCC features and their shifted delta coefficients. In the latter, BNFs are extracted from a multi-lingual phone recognizer neural network. For our experiments, we considered a pre-trained phone recognizer from BUT/phonexia [39] which was trained using 17 Babel languages. For both systems, 100-dimensional i-vectors are extracted using 256 Gaussian mixture components and the obtained i-vectors are transformed by a whitening transformation to be used in the dialect prediction [20], [38].

C. EVALUATION METRICS

The evaluation metrics used are the unweighted average recall (UAR) and accuracy. UAR gives the unbiased scores for the classification and therefore it is considered as the primary evaluation metric for this study. These evaluation metrics were chosen in order not to create any bias towards the majority class as the classes are unevenly distributed.

V. RESULTS

In this section, the results obtained in dialect identification using the proposed system and the baseline systems are reported separately for the STYRIALECT database and the UT-Podcast database.

A. RESULTS FOR THE STYRIALECT DATABASE

In this section, different variants of the SFF spectrogram computation methods described in Section II-B are first investigated to find the best approach for the proposed $\text{BNF}_{\text{Mel-SFF}}$ system in dialect identification. Then, the proposed system with the best approach is compared to the baseline systems and to the fusion system. In both parts, SVM is used as a classifier. Finally, we validate the performance of the proposed system and the fusion system with different classifiers (SVM, MCLR and GLC) in comparison to the best baseline system obtained from the former analysis.

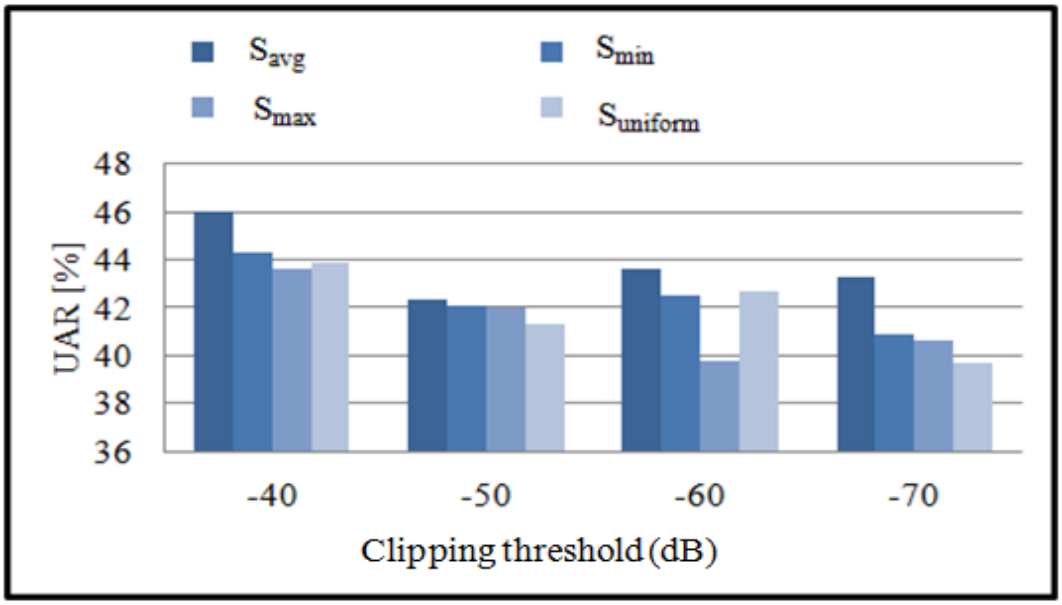


FIGURE 5. Performance (in UAR %) of dialect identification in STYRIALECT using the four spectral operations described in Section II-B (S_{avg}, S_{min}, S_{max}, $S_{uniform}$) and four clipping thresholds (−40 dB, −50 dB, −60 dB, and −70 dB).

In order to find out which of the four spectrogram computation approaches described in Section II-B is best, we conducted dialect identification experiments by using all these approaches and by using four values for the clipping threshold (−40 dB, −50 dB, −60 dB, and −70 dB). The obtained results (in UAR %) are plotted in Figure 5. It can be observed that S_{avg} outperformed the other approaches for all threshold values. Therefore, the feature extraction from the Mel-SFF spectrogram was computed in all further experiments of this study using S_{avg} with a segment duration of 37.5 *ms*.

The dialect identification results reported in UAR [in %] and accuracy [in %] are shown in Table 1 for the proposed $\text{BNF}_{\text{Mel-SFF}}$ method and for the three reference methods by using the SVM classifier. The first column refers to the systems under comparison. In addition to the $\text{BNF}_{\text{Mel-SFF}}$ system and the three reference systems, the table also includes the fusion system (the lowest row of the first column). The second column includes the different clipping thresholds. From the table, it can be observed that the MFCC-based i-vector system performed better than the BUT/phonexia i-vector system. The $\text{BNF}_{\text{Mel-STFT}}$ system showed the best performance among all the systems at the threshold of −70 dB. Among the three baseline systems, the $\text{BNF}_{\text{Mel-STFT}}$ system gave the best performance in both metrics.

The proposed $\text{BNF}_{\text{Mel-SFF}}$ system showed better performance in UAR at the threshold of −40 dB when compared to the $\text{BNF}_{\text{Mel-STFT}}$ baseline system. In addition, the proposed system showed comparable performance with $\text{BNF}_{\text{Mel-STFT}}$ at the threshold levels of −50 dB and −60 dB. Furthermore, the fusion system outperformed the best baseline configuration at the thresholds of −40 dB by 0.42% in UAR.

Furthermore, we compared the dialect identification performance with the three different classifiers described in Section III-D using the $\text{BNF}_{\text{Mel-STFT}}$ baseline system, the proposed $\text{BNF}_{\text{Mel-SFF}}$ system and the fusion system. The results reported in Table 2 are shown for the best configurations from Table 1, i.e., with the threshold of −70 dB for

TABLE 1. Performance in UAR (%) and accuracy of the three baseline systems, the proposed system and the fusion system using the development data of STYRIALECT. SVM is used as the classifier. The utterance length is 2 s.

System	Clipping threshold (dB)	UAR [%]	Accuracy [%]
$\text{i-vector}_{\text{MFCC}}$ system	-	45.0	54.7
$\text{i-vector}_{\text{BUT-BNF}}$ system	-	38.4	52.1
$\text{BNF}_{\text{Mel-STFT}}$ system	-40	43.7	56.4
	-50	44.4	53.3
	-60	44.6	63.9
	-70	**46.7**	66.0
	-80	44.1	56.9
$\text{BNF}_{\text{Mel-SFF}}$ system	-30	45.6	61.0
	-40	**46.0**	60.7
	-50	42.1	57.0
	-60	43.6	56.8
	-70	42.0	56.5
Fusion system ($\text{BNF}_{\text{Mel-STFT}}$ + $\text{BNF}_{\text{Mel-SFF}}$)	-40	**46.9**	57.2
	-50	45.0	52.9
	-60	44.2	61.6
	-70	45.7	57.9

TABLE 2. Performance in UAR (%) in STYRIALECT with different classifiers. The utterance length is 2 s.

System	MCLR	GLC	SVM
$\text{BNF}_{\text{Mel-STFT}}$ system	46.1	42.8	46.7
$\text{BNF}_{\text{Mel-SFF}}$ system	45.4	39.6	46.0
Fusion system	**46.3**	**43.4**	**46.9**

TABLE 3. Performance in UAR (%) in UT-Podcast on the development (dev) and test sets with different classifiers. The utterance length is 17 s.

System	dev			test		
	MCLR	GLC	SVM	MCLR	GLC	SVM
$BNF_{Mel\text{-}STFT}$ system	66.03	50.13	67.39	66.20	54.68	69.62
$BNF_{Mel\text{-}SFF}$ system	71.48	51.12	**73.77**	69.38	57.30	**72.89**
Fusion system	73.38	51.74	**74.78**	71.78	58.90	**74.99**

the $BNF_{Mel\text{-}STFT}$ system and with the threshold of −40 dB for the $BNF_{Mel\text{-}SFF}$ system and for the fusion system. From these experiments, it can be observed that the SVM classifier performed better than the other two classifiers. Furthermore, it can also be observed that the fusion of the BNFs derived from the Mel-STFT and Mel-SFF spectrograms improved the overall performance compared to any of the individual feature extraction methods.

B. RESULTS FOR THE UT-PODCAST DATABASE

From the results reported for the STYRIALECT database in Section V-A, it can be observed that the performance of the i-$vector_{MFCC}$ and i-$vector_{BUT\text{-}BNF}$ systems is poorer compared to the other systems. Hence, these two systems were removed in evaluating the UT-Podcast database. Table 3 presents the UAR results for the three remaining systems separately for the development and test sets and for the three different classifiers. The configurations of these systems are as in Sections III-A and IV-B, except that the clipping threshold is fixed to −40 dB. The autoencoder is trained for 128 epochs with batch size 4 and 128 GRUs in each layer.

From the table, it can be observed that the $BNF_{Mel\text{-}SFF}$ system performed better than the $BNF_{Mel\text{-}STFT}$ system for all classifiers. As in the results discussed in Section V-A, SVM showed higher performance compared to the two other classifiers. The proposed $BNF_{Mel\text{-}SFF}$ system gave a relative improvement of 9.47% and 4.69% in UAR for the development and test set, respectively, when compared to the baseline $BNF_{Mel\text{-}STFT}$ system. The results above support our hypothesis: Since the SFF spectrum is extracted in principle at every sample, the temporal resolution of the spectrogram is preserved. The hidden BNFs derived from the SFF spectrogram showing high spectral and temporal resolution result in better discrimination of speech sounds across dialects.

Furthermore, from the results reported in Tables 2 and 3, it can be observed that the fusion of the BNFs derived from the Mel-STFT and Mel-SFF spectrograms improved the overall performance compared to any of the individual systems. The improvement achieved with the fusion system for both databases shows that there is complementary information between the spectral representations computed by STFT and SFF.

It is to be noted that the results reported in Table 3 were obtained by using the speech sounds of UT-Podcast over the entire length of the utterance (i.e., 17 s). In order to study the effect of the utterance length for dialect identification, additional experiments were carried out using utterance lengths of 10 s and 2 s for the UT-Podcast database. The results reported in Table 4 show that the proposed $BNF_{Mel\text{-}SFF}$ system performed consistently for all utterance lengths. Furthermore, it can be observed that the proposed system showed a clearly larger improvement (15.35% relative) compared to the STFT-based reference system for shorter utterances than for longer utterances (4.69% relative). Furthermore, the fusion system showed an improvement for both utterance lengths compared to the individual reference systems, again indicating complementary information between the features.

TABLE 4. Performance in UAR (%) in UT-Podcast using the development (dev) and test sets with different utterance lengths.

System	length (s)	dev	test
$BNF_{Mel\text{-}STFT}$ system	10	63.15	63.09
$BNF_{Mel\text{-}SFF}$ system	10	**70.12**	**69.96**
Fusion system	10	**70.54**	**72.32**
$BNF_{Mel\text{-}STFT}$ system	2	54.01	55.50
$BNF_{Mel\text{-}SFF}$ system	2	**61.30**	**64.02**
Fusion system	2	**65.74**	**64.58**

VI. SUMMARY AND CONCLUSION

This study explored the use of the Mel-weighted single frequency filtering spectrogram for dialect identification using the STYRIALECT and UT-Podcast databases. Dialects were identified by training an autoencoder with the Mel-SFF spectrogram and by feeding the bottleneck features of the autoencoder to a classifier. The proposed Mel-SFF spectrogram gave better performance compared to the i-vector based baseline systems. Furthermore, the fusion of the unsupervised representations (BNFs) computed from the Mel-SFF and Mel-STFT spectrograms using the sequence-to-sequence autoencoders yielded the best UAR score (46.9%) for the STYRIALECT database. In UT-Podcast, the proposed and fusion systems gave a relative UAR improvement of 4.69% and 7.71% compared to the Mel-STFT spectrogram-based baseline system, respectively. Furthermore, the proposed system showed better performance especially in short utterances compared to the baseline system in the experiments with the UT-Podcast data. Therefore, we conclude that the high spectral and temporal resolution of the SFF spectrum leads to an improvement in dialect identification for the studied

German and English dialects. In addition, we conclude that the proposed Mel-SFF spectrogram system distinguishes dialects better from short utterances than its STFT-based reference system. In the future, we plan to explore the Mel-SFF spectrogram derived features for dialect identification in noisy conditions [27], [28], [30] and for larger corpora.

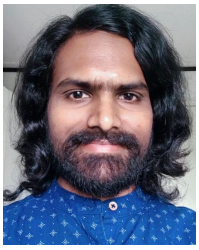
ACKNOWLEDGMENT

Rashmi Kethireddy would like to thank the University Grants Commission, India, for supporting the Ph.D. degree. Sudarsana Reddy Kadiri would like to thank the Academy of Finland for supporting his stay in Finland as a Postdoctoral Researcher.

REFERENCES

[1] M. Tjalve and M. Huckvale, "Pronunciation variation modelling using accent features," in *Proc. Interspeech*, vol. 2005, pp. 1341–1344.
[2] P. Motlicek, P. N. Garner, N. Kim, and J. Cho, "Accent adaptation using subspace Gaussian mixture models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 7170–7174.
[3] F. Biadsy, "Automatic dialect and accent recognition and its application to speech recognition," Ph.D. dissertation, Graduate School Arts Sci., Columbia Univ., New York, NY, USA, 2011.
[4] Y. Lei and J. H. L. Hansen, "Dialect classification via text-independent training and testing for Arabic, Spanish, and Chinese," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 19, no. 1, pp. 85–96, Jan. 2011.
[5] M. Najafian, S. Safavi, P. Weber, and M. Russell, "Identification of British English regional accents using fusion of i-vector and multi-accent phonotactic systems," in *Proc. Odyssey*, Jun. 2016, pp. 132–139.
[6] F. Biadsy, J. Hirschberg, and N. Habash, "Spoken Arabic dialect identification using phonotactic modeling," in *Proc. EACL Workshop Comput. Approaches Semitic Lang. Semitic*, 2009, pp. 53–61.
[7] M. A. Zissman, T. P. Gleason, D. M. Rekart, and B. L. Losiewicz, "Automatic dialect identification of extemporaneous conversational, latin American Spanish speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. Conf.*, May 1996, pp. 777–780.
[8] F. S. Richardson, W. M. Campbell, and P. A. Torres-Carrasquillo, "Discriminative n-gram selection for dialect recognition," in *Proc. Interspeech*, 2009, pp. 192–195.
[9] R. Huang, J. H. L. Hansen, and P. Angkititrakul, "Dialect/accent classification using unrestricted audio," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 15, no. 2, pp. 453–464, Feb. 2007.
[10] J. H. Hansen and L. M. Arslan, "Foreign accent classification using source generator based prosodic features," *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, vol. 1, pp. 836–839, 1995.
[11] L. M. Arslan and J. H. L. Hansen, "A study of temporal features and frequency characteristics in American English foreign accent," *J. Acoust. Soc. Amer.*, vol. 102, no. 1, pp. 28–40, Jul. 1997.
[12] L. Wai Kat and P. Fung, "Fast accent identification and accented speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 1999, pp. 221–224.
[13] M. Huckvale, "ACCDIST: A metric for comparing speakers' accents," in *Proc. Interspeech*, 2004, pp. 29–32.
[14] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. D. Jr, "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features," in *Proc. Int. Conf. Spoken Lang. Process. (INTERSPEECH)*, 2002.
[15] S. Safavi, A. Hanani, M. Russell, P. Jancovic, and M. J. Carey, "Contrasting the effects of different frequency bands on speaker and accent identification," *IEEE Signal Process. Lett.*, vol. 19, no. 12, pp. 829–832, Dec. 2012.
[16] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digit. Signal Process.*, vol. 10, nos. 1–3, pp. 19–41, Jan. 2000.
[17] F. Biadsy, J. Hirschberg, and D. P. W. Ellis, "Dialect and accent recognition using phonetic-segmentation supervectors," in *Proc. Interspeech*, 2011, pp. 745–748.
[18] H. Behravan, V. Hautamaki, S. M. Siniscalchi, T. Kinnunen, and C.-H. Lee, "I-vector modeling of speech attributes for automatic foreign accent recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 1, pp. 29–41, Jan. 2016.
[19] A. Hanani, M. J. Russell, and M. J. Carey, "Human and computer recognition of regional accents and ethnic groups from British English speech," *Comput. Speech Lang.*, vol. 27, no. 1, pp. 59–74, Jan. 2013.
[20] H. Behravan, V. Hautamäki, and T. Kinnunen, "Foreign accent detection from spoken Finnish using i-vectors," in *Proc. Interspeech*, 2013, pp. 79–83.
[21] A. DeMarco and S. J. Cox, "Iterative classification of regional British accents in i-vector space," in *Proc. Symp. Mach. Learn. Speech Lang. Process.*, 2012, pp. 1–4.
[22] Y. Song, R. Cui, X. Hong, I. Mcloughlin, J. Shi, and L. Dai, "Improved language identification using deep bottleneck network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 4200–4204.
[23] P. Matejka, L. Zhang, T. Ng, O. Glembek, J. Z. Ma, B. Zhang, and S. H. Mallidi, "Neural network bottleneck features for language identification," in *Proc. ODYSSEY*, 2014, pp. 1–6.
[24] Q. Zhang and J. H. L. Hansen, "Language/dialect recognition based on unsupervised deep learning," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 5, pp. 873–882, May 2018.
[25] B. W. Schuller, A. Batliner, C. Bergler, F. B. Pokorny, J. Krajewski, M. Cychosz, R. Vollmann, S.-D. Roelen, S. Schnieder, E. Bergelson, A. Cristia, A. Seidl, A. S. Warlaumont, L. Yankowitz, E. Nöth, S. Amiriparian, S. Hantke, and M. Schmitt, "The INTERSPEECH 2019 computational paralinguistics challenge: Styrian dialects, continuous sleepiness, baby sounds & orca activity," in *Proc. Interspeech*, Sep. 2019, pp. 2378–2382.
[26] A. Shahin, F. Michael, C. Nicholas, and S. Björn, "Sequence to sequence autoencoders for unsupervised representation learning from audio," in *Proc. Detection Classification Acoustic Scenes Events*, 2017, pp. 17–21.
[27] V. Pannala, G. Aneeja, S. R. Kadiri, and B. Yegnanarayana, "Robust estimation of fundamental frequency using single frequency filtering approach," in *Proc. Interspeech*, Sep. 2016, pp. 2155–2159.
[28] S. R. Kadiri and B. Yegnanarayana, "Epoch extraction from emotional speech using single frequency filtering approach," *Speech Commun.*, vol. 86, pp. 52–63, Feb. 2017.
[29] B. T. Nellore, R. Prasad, S. R. Kadiri, S. V. Gangashetty, and B. Yegnanarayana, "Locating burst onsets using SFF envelope and phase information," in *Proc. Interspeech*, Aug. 2017, pp. 3023–3027.
[30] G. Aneeja, S. R. Kadiri, and B. Yegnanarayana, "Detection of glottal closure instants in degraded speech using single frequency filtering analysis," in *Proc. Interspeech*, Sep. 2018, pp. 2300–2304.
[31] J. H. L. Hansen, S. S. Gray, and W. Kim, "Automatic voice onset time detection for unvoiced stops (/p/,/t/,/k/) with application to accent classification," *Speech Commun.*, vol. 52, no. 10, pp. 777–789, Oct. 2010.
[32] G. Aneeja and B. Yegnanarayana, "Single frequency filtering approach for discriminating speech and nonspeech," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 4, pp. 705–717, Apr. 2015.
[33] S. R. Kadiri and B. Yegnanarayana, "Analysis of aperiodicity in artistic noh singing voice using an impulse sequence representation of excitation source," *J. Acoust. Soc. Amer.*, vol. 146, no. 6, pp. 4446–4457, Dec. 2019.
[34] M. Freitag, S. Amiriparian, S. Pugachevskiy, N. Cummins, and B. Schuller, "AuDeep: Unsupervised learning of representations from audio with deep recurrent neural networks," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 6340–6344, 2017.
[35] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.
[36] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, and J. Vanderplas, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.
[37] J. H. L. Hansen and G. Liu, "Unsupervised accent classification for deep data fusion of accent and language information," *Speech Commun.*, vol. 78, pp. 19–33, Apr. 2016.
[38] A. Abad, E. Ribeiro, F. Kepler, R. Astudillo, and I. Trancoso, "Exploiting phone log-likelihood ratio features for the detection of the native language of non-native English speakers," in *Proc. Interspeech*, Sep. 2016, pp. 2413–2417.
[39] A. Silnova, P. Matejka, O. Glembek, O. Plchot, O. Novotny, F. Grezl, P. Schwarz, L. Burget, and J. Cernocky, "BUT/Phonexia bottleneck feature extractor," in *Proc. Odyssey Speaker Lang. Recognit. Workshop*, Jun. 2018, pp. 283–287.

RASHMI KETHIREDDY (Graduate Student Member, IEEE) received the Bachelor of Technology degree in information technology (I.T.) from the Kakatiya Institute of Technology and Science, Warangal, India, in 2011, and the Master of Technology degree in computer science engineering from Osmania University, Hyderabad, India, in 2017. She was with IT services for a period of two years. She is currently a Ph.D. Scholar with the International Institute of Information Technology-Hyderabad (IIIT-H). Her research interests include speech signal processing, acoustic analysis, machine learning, speech dialectal challenges, and speech dialect identification. She qualified the University Grant Commission National Eligibility Test (UGC-NET). She also received the Junior Research Fellowship (JRF).



SUDARSANA REDDY KADIRI (Member, IEEE) received the Bachelor of Technology degree in electronics and communication engineering (ECE) from Jawaharlal Nehru Technological University (JNTU), Hyderabad, India, in 2011, the M.S. degree in research during 2011–2014, and later converted to the Ph.D. degree, and received the Ph.D. degree from the Department of ECE, International Institute of Information Technology, Hyderabad (IIIT-H), India, in 2018. He was a Teaching Assistant of several courses with IIIT-H from 2012 to 2018. He is currently a Postdoctoral Researcher with the Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland. His research interests include signal processing, speech analysis, speech synthesis, paralinguistics, affective computing, voice pathologies, machine learning, and auditory neuroscience.

PAAVO ALKU (Fellow, IEEE) received the M.Sc., Lic.Tech., and Dr.Sc. (Tech.) degrees from the Helsinki University of Technology, Espoo, Finland, in 1986, 1988, and 1992, respectively. He was an Assistant Professor with the Asian Institute of Technology, Bangkok, Thailand, in 1993. He was an Assistant Professor and a Professor with the University of Turku, Finland, from 1994 to 1999. He was also an Academy Professor with the Academy of Finland from 2015 to 2019. He is currently a Professor of speech communication technology with Aalto University, Espoo. He has published more than 200 peer-reviewed journal articles and more than 200 peer-reviewed conference papers. His research interests include analysis and parameterization of speech production, statistical parametric speech synthesis, spectral modeling of speech, speech-based biomarking of human health, and cerebral processing of speech. He is a Fellow of ISCA. He serves as an Associate Editor for *Journal of the Acoustical Society of America*.

SURYAKANTH V. GANGASHETTY (Member, IEEE) received the Ph.D. degree in neural network models for recognition of consonant-vowel units of speech in multiple languages from IIT Madras, in 2005. He was a Senior Project Officer with the Speech and Vision Laboratory, IIT Madras. He was a Faculty Member with BIET Davangere from 1991 to 1999. He was also a Visiting Research Scholar with OGI Portland, USA, for a period of three months, in Summer 2001. He has been a Faculty Member with IIIT-H, India, since August 2006. He held a postdoctoral position (PDF) with Carnegie Mellon University (CMU), Pittsburgh, PA, USA, from April 2007 to July 2008. He is the author of about 150 papers published in national and international journals, conferences, and edited volumes. His research interests include speech processing, neural networks, machine learning, natural language processing, and artificial intelligence. He is a Life Member of CSI, IE, IUPRAI, ASI, IETE, ORSI, and ISTE. He served as the Local Organizing Chair for the INTERSPEECH-2018 Conference, Hyderabad, India, in September 2018. He has reviewed papers for reputed journals and conferences.