Wavelet Scattering Transform for Improving Generalization in Low-Resourced Spoken Language Identification

Spandan Dey, Premjeet Singh, Goutam Saha

Department of Electronics & Electrical Communication Engineering,
Indian Institute of Technology Kharagpur, West Bengal, India

sd21@iitkgp.ac.in, premsingh@iitkgp.ac.in, gsaha@ece.iitkgp.ac.in

Abstract

Commonly used features in spoken language identification (LID), such as mel-spectrogram or MFCC, lose high-frequency information due to windowing. The loss further increases for longer temporal contexts. To improve generalization of the low-resourced LID systems, we investigate an alternate feature representation, wavelet scattering transform (WST), that compensates for the shortcomings. To our knowledge, WST is not explored earlier in LID tasks. We first optimize WST features for multiple South Asian LID corpora. We show that LID requires low octave resolution and frequency-scattering is not useful. Further, cross-corpora evaluations show that the optimal WST hyper-parameters depend on both train and test corpora. Hence, we develop fused ECAPA-TDNN based LID systems with different sets of WST hyper-parameters to improve generalization for unknown data. Compared to MFCC, EER is reduced upto 14.05% and 6.40% for same-corpora and blind VoxLingua107 evaluations, respectively.

Index Terms: Language identification, Cross-corpora evaluation, Wavelet scattering transform, ECAPA-TDNN.

1. Introduction

In the context of spoken language identification (LID), the term *low-resourced* indicates the lack of large-scale multilingual speech corpora with verified ground truths. Developing effective low-resourced LID systems is important for multilingual human-to-computer interaction applications for the global population [1]. Often, the generalization of the low-resourced LID systems is challenged due to training with a small in-house developed corpus, which lacks diversities in non-lingual characteristics [2]. Towards this, we aim to improve the generalization of low-resourced LID systems by applying feature representations that are robust against non-lingual variations across multiple corpora. To assess generalization, we follow cross-corpora evaluation protocols, which are particularly useful for low-resourced scenarios [3].

Mel-spectrograms and mel frequency cepstral coefficients (MFCCs), are one of the most widely used features in the LID literature. Both these features were originally developed for automatic speech recognition (ASR). By applying mel filterbanks on short-time Fourier transform (STFT) representations, these features generate features resembling the output of the cochlea [4]. The mel-based features further provide stability toward local time translation and deformation, usually up to a window of 25ms [5]. This temporal span works well for phoneme recognition purposes in ASR but may not be the most suitable choice for recognizing languages, which requires longer phonotactic information as well [6]. One of the key issues with these features is the loss of information due to the time-domain convolution with a low-pass filter due to windowing [5]. For classification tasks requiring a larger temporal context, capturing these features over a longer temporal window incurs even more information loss.

Hence, in this work, we introduce alternative representations in the LID task, that provide stability against deformations and reduce information loss even with longer temporal contexts. Mallat proposed *wavelet scattering transform* (WST) extending the MFCCs by computing modulation spectrum with the cascaded application of wavelet filter banks and modulus non-linearities [7]. Abiding by the *Lipschitz continuity* condition, WST extracts stable representations for variations due to time-shifts and time-warping deformations over a larger temporal span without losing the high-frequency information [5]. In [5], the authors applied WST for music genre and phoneme classification. Environmental sound classification was performed utilizing WST in [8, 9, 10]. Bruna et al. [11] used scattering moments to synthesize audio textures. Joy et al. [12] applied scattering power spectrum for speech recognition. In [13] and [14], WST was applied for music processing applications. Recently, the potentials of WST are explored in speech emotion recognition [15] and speaker recognition tasks [16].

However, to our knowledge, this is the first study applying scattering networks in a LID task. Therefore, we first systematically formulate the work, starting by exploring the fundamental questions: **(Q1)** What are the WST hyper-parameters suitable for the LID tasks? **(Q2)** Are the optimized hyper-parameters corpora dependent? **(Q3)** How much performance improvements can we obtain from the conventional MFCCs by optimizing the WST hyper-parameters? **(Q4)** Is scattering transform across the log-frequency dimension useful for LID? After answering these fundamentals, we then focus on improving cross-corpora LID generalization with WST features. The answer to Q2 reveals the dependency of optimal WST hyper-parameters on both training and evaluation data. Hence, concerning unknown real-world test conditions, we develop multi-WST fusion based LID systems encompassing the representations generated with different WST hyper-parameters.

2. Methodology

2.1. Shortcomings of Fourier-based representations

Consider a signal $x(t)$ with its Fourier transform (FT) denoted by $\widehat{x}(\omega)$. For a time-shift c expressed by $x_c(t) = x(t - c)$, the corresponding FT is $\widehat{x}_c(\omega) = e^{-i\omega c}\widehat{x}(\omega)$. The modulus of FT removes the additional phase part $e^{-i\omega c}$ and makes the representation translation invariant, $|\widehat{x}_c(\omega)| = |\widehat{x}(\omega)|$. Let the short-time Fourier transform (STFT) or spectrogram of $x(t)$ is defined as $|\widehat{x}(t,\omega)| = \left| \int x(u)\,\phi(u - t)\,e^{-i\omega u}\,du \right|$. Here, ϕ is a window with duration T. For $c << T$, STFT is local time-

shift invariance. However, STFT is not stable to time-warping deformations, which often takes place with audio data [5]. Here, the notion of stability is defined by the Lipschitz continuity condition, which states that a transformed representation $\Phi(x)$ is stable to deformation by amount $\sup_t |\tau'(t)|$ if,

$$||\Phi(x) - \Phi(x_\tau)|| \leq C \sup_t |\tau'(t)| \, ||x|| \, . \quad (1)$$

Here, $C > 0$ is a constant and measures the stability. Considering a deformation, $\tau(t) = \epsilon t$, where $0 < \epsilon \ll 1$ the Fourier transform of $x_\tau(t) = x(t - \tau(t)) = x((1 - \epsilon)t)$, for $|\tau'(t)| < 1$, is $\widehat{x}_\tau(\omega) = (1 - \epsilon)^{-1} \widehat{x}((1 - \epsilon)^{-1}\omega)$. Hence, time deformation leads to frequency translation of $\epsilon|\omega|$. Following Eq. 1, we obtain $|| \, |\widehat{x}| - |\widehat{x_\tau}| \, || \leq C \epsilon ||x||$. This implies that Lipschitz continuity can be violated by spectrograms for higher values of ω.

To impose stability against time-warping deformation, averaging with mel filters ($\widehat{\psi}_\lambda(\omega)$), with center frequency λ is done over spectrograms, i.e.,

$$\mathrm{M}x(t, \lambda) = \frac{1}{2\pi} \int |\widehat{x}(t, \omega)|^2 \, |\widehat{\psi}_\lambda(\omega)|^2 d\omega \, , \quad (2)$$

where, $\widehat{x}(t, \omega)$ is the FT of $x_t(u) = x(u)\phi(u - t)$. The time-domain equivalent of Eq. 2 is given as,

$$\mathrm{M}x(t, \lambda) = \int \left| \int x(u)\phi(u - t)\psi_\lambda(v - u)du \right|^2 dv \quad (3)$$

Now, if the length (T) of window ϕ is much larger than the temporal support of $\psi_\lambda(t)$, we can consider $\phi(t)$ as constant over the span of $\psi_\lambda(t)$. Hence, considering $\phi(u - t)\psi_\lambda(v - u) \approx \phi(v - t)\psi_\lambda(v - u)$ in Eq. 3,

$$\mathrm{M}x(t, \lambda) \approx \int \left| \int x(u)\psi_\lambda(v - u)du \right|^2 |\phi(v - t)|^2 dv \quad (4)$$

$$= |x \star \psi_\lambda|^2 \star |\phi|^2(t) \, . \quad (5)$$

The mel-filters $\widehat{\psi}_\lambda(\omega)$ with center frequency λ have constant-Q bandwidths (BW) at high frequencies. So, the BW of $\widehat{\psi}_\lambda(\omega)$ is λ/Q, which is sufficiently large at higher λs to encompass stability to time-warping deformations. However, as evident from Eq. 5, the windowing performed on the signal is equivalent to time averaging of spectrograms by the low-pass filter $\phi(t)$. Hence, there is an inherent downside in mel-spectrograms/MFCCs of high-frequency information loss. Instead of the standard window size, usually fixed to $25\,\mathrm{ms}$, if some classification task demands a higher temporal context, the risk of higher information loss restricts the usefulness of mel-spectrograms/MFCCs. Hence, to restore the high-frequency information while maintaining stability to deformations over a longer span, wavelet scattering transform can be used.

2.2. Wavelet scattering transform

Scattering transform applies cascade of wavelet transforms, with constant Q-factor wavelet filters and modulus operators for restoring high-frequency information lost due to averaging. Consider, for $\lambda > 0$ a dilated wavelet band pass filter $\psi_\lambda(t) = \lambda\psi(\lambda t)$ with frequency-domain representation $\widehat{\psi}_\lambda(\omega) = \widehat{\psi}\left(\frac{\omega}{\lambda}\right)$. The center frequency of $\widehat{\psi}_\lambda(\omega)$ is λ (normalized) and BW is λ/Q, with Q denoting the octave resolution of wavelet filters. Hence, $\lambda = 2^{k/Q}$ with $k \in \mathbb{Z}$. The wavelet filters span the entire frequency range of the input signal. Each filter $\psi_\lambda(t)$ has a temporal span of $2\pi Q/\lambda$. To ensure this span is less than T, λs are only defined for $\lambda \geq 2\pi Q/T$. For lower frequencies $[0, 2\pi Q/T]$, $Q - 1$ equally spaced filters with BW

$2\pi/T$ are designed. The wavelet transform of a signal x is expressed as:

$$Wx = \left(x \star \phi(t) \, , \, x \star \psi_\lambda(t) \right)_{t \in \mathbb{R}, \lambda \in \Lambda} \, . \quad (6)$$

Here, ϕ is the low-pass filter with BW $2\pi/T$ and the set of all higher ($\geq 2\pi Q/T$) center frequencies are denoted by Λ. For translation invariance, as a contractive non-linear operator [7], modulus operation is applied over Wx:

$$|W_1|x = \Big(\underbrace{x \star \phi(t)}_{S_0 x(t)}, \, \underbrace{|x \star \psi_{\lambda_1}(t)|}_{U_1 x(t, \lambda_1)} \Big)_{t \in \mathbb{R}, \lambda_1 \in \Lambda_1} \, . \quad (7)$$

The first stage of the wavelet transform applies wavelets with center frequencies Λ_1 and resolution Q_1. From Eq. 7, we set $S_0 x(t) = x \star \phi(t)$, which is locally translation invariant due to averaging with ϕ. The term $|x \star \psi_{\lambda_1}(t)|$ provides a time-frequency representation of x where the varying bandwidth filters ψ_{λ_1} introduce the required deformation stability [5]. The resultant representation is then low-pass filtered with ϕ so as to capture long temporal context and is finally denoted as the *first-order scattering coefficients* $S_1 x(t, \lambda_1) = |x \star \psi_{\lambda_1}| \star \phi(t)$.

$S_1 x(t, \lambda_1)$, also known as *scalogram*, approximates mel-spectrograms if $Q_1 = 8$ [5]. The information lost in the scalogram due to low-pass filtering is further retrieved by applying the second stage of wavelet filters:

$$|W_2| \, |x \star \psi_{\lambda_1}| = \Big(\underbrace{|x \star \psi_{\lambda_1}| \star \phi}_{S_1 x(t, \lambda_1)}, \, \underbrace{||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}|}_{U_2 x(t, \lambda_1, \lambda_2)} \Big)_{\lambda_2 \in \Lambda_2} \, . \quad (8)$$

From Eq. 8, we compute the *second-order scattering coefficients* (also referred as modulation spectrum) $S_2 x(t, \lambda_1, \lambda_2) = ||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi(t) = U_2 x(t, \lambda_1, \lambda_2) \star \phi(t)$.

The second layer coefficients capture the information lost in first layer over longer temporal context of ϕ. Usually, we follow $Q_2 = 1$ to capture the short-spanned transients and to generate sparse representation [5]. Iteratively, we can apply successive stages of modulus wavelet transform followed by low-pass filtering to generate the higher-order scattering coefficients. For decorrelation and to create invariance against multiplicative factors, at each layer, scattering coefficients are log-normalized by the previous layer's coefficients [5]. Algorithm 1 summarizes the process of m-th order WST feature extraction.

2.3. WST for frequency transposition invariance

The WST can further be operated along the log-frequency ($log\lambda$) axis of the scalograms for frequency transposition invariance. Frequency transposition is characterized by the inter-speaker translation of spectral components in the log-frequency scale, affecting the pitch and spectral envelope information [5]. The frequency-domain WST coefficient computation is similar to time-domain WST, with t replaced by $log\lambda$. For LID tasks, where robustness against speaker variability is important, the exploration of frequency-domain scattering is interesting.

3. Experiment details

3.1. Database description

We use three most widely used South Asian LID corpora, IIITH-ILSC [17] (IIITH), LDC 2017S14 [18] (LDC), and IITKGP-MLILSC [19] (KGP) with five languages, *Bengali*, *Hindi*, *Punjabi*, *Tamil*, and *Urdu*. IIITH and KGP data are pre-partitioned in speaker disjoint train and test sets. We manually split the LDC data into speaker disjoint train and test sets

Algorithm 1 Computing WST features up to order $m > 1$.

Input: $x(t)$, $Q_1, Q_2, \cdots, Q_m$, ϕ

Output: $\widetilde{S}x$

1: **procedure** (Applying cascaded modulus wavelet transform $|W|$ followed by averaging with ϕ)
2: $\quad U_0 x = x$
3: $\quad$ **for** $l = 1 : m$ **do**
4: $\quad\quad |W_{l+1}| U_l x = (S_l x, U_{l+1} x)$
5: $\quad\quad$ **if** $l == 1$ **then**
6: $\quad\quad\quad \widetilde{S}_1 x(t, \lambda_1) = log\left(\frac{S_1 x(t, \lambda_1)}{|x| \star \phi(t) + \epsilon}\right)$ $\quad \triangleright$ Log-normalization
7: $\quad\quad$ **else**
8: $\quad\quad\quad \widetilde{S}_l x(t, \lambda_1, \cdots, \lambda_l) = log\left(\frac{S_l x(t, \lambda_1, \cdots, \lambda_l)}{S_{l-1} x(t, \lambda_1, \cdots, \lambda_{l-1}) + \epsilon}\right)$
9: $\quad\quad$ **end if**
10: $\quad$ **end for**
11: $\quad \widetilde{S}x = \left(\widetilde{S}_0 x(t), \widetilde{S}_1 x(t, \lambda_1), \cdots, \widetilde{S}_m x(t, \lambda_1, \lambda_2, \cdots, \lambda_m)\right)$
12: **end procedure**

---

following $80 : 20$ ratio. The three train sets are further split into speaker disjoint train and validation sets using the same ratio. The train set is further augmented and then sampled to make it two folds following the augmentation procedure followed in [20], which applied babble, music, noise samples, and different room impulse responses as perturbations to the utterances. All the utterances are re-sampled to $8\,\text{kHz}$ and split into $3\,\text{s}$ chunks. With the training and validation sets of each corpus, we train three standalone LID systems. During the evaluation, with the test sets of all three corpora, we perform same-corpora and cross-corpora evaluations. We also use subset of the VoxLingua107 corpus [21] as a blind evaluation set, entirely unknown during the system development stages.

3.2. Data pre-processing and feature extraction

The audio signals are first processed with an energy-based voice activity detector (VAD) to discard the silence-detected frames. Then, we extract WST features with different hyper-parameter sets, which include the temporal span (T) of ϕ, the order of WST coefficients (m) indicating the number of layers, and octave resolutions ($Q = [Q_1, Q_2, \cdots, Q_m]$) at each layer. We use Morlet wavelet and explore it for $T = [256, 512, \cdots, 16384]$. With the sampling rate of $8\,\text{kHz}$, it approximately covers the temporal span from $30\,\text{ms}$ to $2\,\text{s}$. However, for all three corpora, we found a prominent drop in LID performance for $T > 2048$. Hence, we consider LID performances up to $T = 2048$. For this range of T, the signal energy contained by the 3-rd WST layer onwards becomes gradually negligible ($< 1\%$). Hence, following [5], we set $m = 2$. For Q_1, we use values 2, 4, and 8. Following the literature [5, 12, 15], to capture the finer temporal transients, we set $Q_2 = 1$. Similarly, for the frequency-domain WST, following the literature conventions [5, 15], we set $m = 1$ and use octave resolutions (Q_f) between 3 to 8. Following Algorithm 1, the time-domain WST features from the 0-th, 1-st, and 2-nd layers are concatenated after log-normalization. The frequency-domain WST features are appended with the time-domain WST features for LID training and are finally processed with cepstral mean subtraction (CMS). Following the South Asian LID literature [22, 23, 3, 24] as baseline reference, we also train the LID systems using 20-dimensional MFCCs with $25\,\text{ms}$ window, $10\,\text{ms}$ hop-length, 20 mel-filters, and processed with CMS. Following NIST LRE and OLR challenge protocols [25, 26], we use equal error rate (EER) and C_{avg} as performance metrics.

3.3. Classifier description

From a computation perspective, WST is similar to the convolutional neural network (CNN) architecture, while the filters are pre-defined, not learned [7]. Hence, after the hand-crafted convolution layers, we use time-delay neural network (TDNN) based stacks of dilated CNN layers. We use the ECAPA-TDNN [27] architecture to train the LID models[1]. ECAPA-TDNN extends the x-vector architecture [20] by replacing the frame-level TDNN layers with squeeze-excitation-based residual blocks (SE-Res2), multi-layer feature aggregation, and a channel attentive pooling layer. In different speech processing tasks, this architecture is reported to outperform several other TDNN-based models [24, 27, 28]. The classifiers are trained end-to-end using 30 epochs and batch size 64. AdamW optimizer is used with additive margin (AM) softmax loss [29]. The learning rate (LR) is 0.001 following a reduce-on-plateau based LR scheduler with patience of 5 and scale 0.1.

4. Results & discussions

4.1. WST hyper-parameters and LID performances

We first extract WST features with different hyper-parameter sets by varying T and Q_1 (as mentioned in Section 3) and train LID systems using the training and validation data of each corpus. The corresponding same-corpora evaluation performances are presented in Table 1. We use EER values to denote the best LID performances, which are used to find out the best hyper-parameter set for each corpus. The best performing hyper-parameter set for IIITH corpus (I_{hp}) is for $T = 256$ and $Q_1 = 2$ (denoted as Q in Table 1). Similarly, for LDC we obtain L_{hp} for $T = 1024$ and $Q = 2$. For the KGP corpus, the optimal hyper-parameter (K_{hp}) is the same as I_{hp}. Two key observations from Table 1 are: (i) all three corpora show the best LID performance for $Q = 2$, indicating that highly localized frequency cues are not very crucial for LID tasks. (ii) IIITH and KGP both contain broadcast news reads and has optimal $T = 256$. Whereas LDC contains conversational telephone speech (CTS) and has a higher optimal $T = 1024$. This observation indicates the requirement for a larger temporal context for LID in spontaneous conversations. To illustrate how the WST information is useful, in Fig. 1, we plot the modulation spectrums averaged across the IIITH training utterances. The plots show distinct 3D surface patterns for different languages, indicating their efficacy in the LID tasks.

4.2. Impact of frequency-domain WST

For each corpus, we extend the best-performing time-domain WST hyper-parameters by using frequency scattering (f-WST) with Q_f varying between 3 to 8. With the f-WST feature, we train LID systems and report their performances in Fig. 2. For comparison, we also present the best-performing time-domain WST system's LID performances and show that f-WST does not improve performance. Hence, we only consider the time-domain WST features in the subsequent experiments. We also find that KGP corpus, with the lowest number of speakers among the three corpora, exhibits the highest EER improvement by varying Q_f due to invariance for speaker variations by f-WST. While IIITH, already having much larger speakers, is inherently speaker-robust, and f-WST does not help much.

---

[1] https://github.com/Snowdar/asv-subtools

Table 1: *Impact of different WST hyper-parameters on LID performances (EER (%) / $C_{\mathrm{avg}} * 100$) using three LID corpora.*

Corpus	Baseline MFCC	T = 256			T = 512			T = 1024			T = 2048		
		Q=2	Q=4	Q=8	Q=2	Q=4	Q=8	Q=2	Q=4	Q=8	Q=2	Q=4	Q=8
IIITH	9.74 / 11.51	**7.53 / 8.20**	7.56 / **8.14**	7.75 / 8.87	8.35 / 9.71	7.92 / 9.11	10.73 / 11.41	12.58 / 14.16	12.69 / 13.96	12.58 / 14.16	15.25 / 15.79	19.16 / 20.32	20.66 / 20.72
LDC	21.86 / 25.70	17.10 / 19.99	17.64 / 20.41	18.70 / 21.88	13.24 / 15.80	14.38 / 16.53	13.86 / 16.63	**12.87 / 15.41**	14.61 / 17.74	13.53 / 16.09	14.99 / 16.51	15.59 / 17.73	15.48 / 18.57
KGP	12.36 / 8.75	**5.55 / 5.83**	8.21 / 7.51	8.10 / 7.91	8.00 / 7.75	8.12 / 8.12	7.00 / 6.62	9.75 / 9.00	13.00 / 13.00	9.00 / 8.62	19.00 / 18.37	15.12 / 15.12	22.00 / 21.62

Table 2: *Cross-corpora LID performances (EER (%) / $C_{\mathrm{avg}} * 100$) using IIITH, LDC, and KGP with different WST hyper-parameters.*

| Train corpus	Test corpus	Baseline MFCC	T = 256			T = 512			T = 1024			T = 2048		
			Q=2	Q=4	Q=8	Q=2	Q=4	Q=8	Q=2	Q=4	Q=8	Q=2	Q=4	Q=8
IIITH	LDC	42.43 / 44.82	41.72 / 39.26	43.73 / 41.91	43.91 / 41.51	**35.17 / 36.77**	37.94 / 40.25	39.47 / 42.78	40.29 / 43.74	35.50 / 40.12	40.29 / 43.74	36.30 / 39.64	38.13 / 43.94	38.16 / 43.07
	KGP	**34.83 / 32.62**	36.11 / 33.50	43.51 / 38.22	51.85 / 40.82	41.00 / 34.00	40.75 / 32.87	44.00 / 41.62	45.75 / 44.37	42.00 / 40.12	45.75 / 44.37	48.00 / 46.75	52.75 / 47.87	49.87 / 44.12
LDC	IIITH	46.35 / 43.21	38.49 / 39.10	33.67 / 32.86	37.62 / 35.05	40.72 / 37.09	37.45 / 35.42	**31.86 / 31.95**	38.69 / 34.99	37.73 / 35.61	40.71 / 37.16	39.68 / 37.69	38.12 / 36.80	41.42 / 40.33
	KGP	42.95 / 39.59	40.74 / 43.05	45.37 / 44.86	46.29 / 42.34	40.00 / 39.00	37.00 / **33.12**	**36.00** / 35.50	45.00 / 42.62	43.00 / 39.50	38.00 / 39.00	48.87 / 47.25	49.00 / 43.75	44.00 / 40.12
KGP	IIITH	**36.52 / 33.59**	43.12 / 46.96	41.70 / 40.62	49.13 / 44.27	41.33 / 41.95	44.81 / 41.14	45.38 / 40.39	41.00 / 40.00	42.85 / 39.32	41.92 / 41.66	41.88 / 43.18	47.72 / 47.48	46.33 / 44.88
	LDC	47.25 / 45.49	51.28 / 44.61	**44.91** / 41.14	47.91 / 42.58	46.33 / 41.66	45.30 / 41.23	45.33 / 41.27	46.89 / 42.43	46.22 / 41.11	45.03 / 41.51	45.09 / 41.15	46.34 / **40.70**	48.45 / 42.19
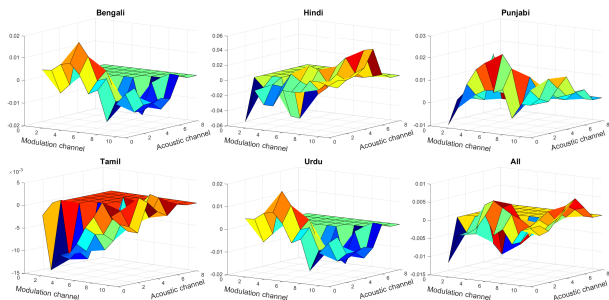
Figure 1: *Visualization of modulation spectrum for each acoustic channel, averaged over the IIITH training utterances.*

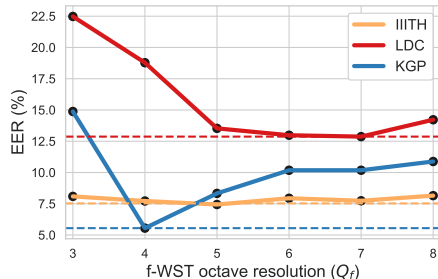Figure 2: *Impact of f-WST octave resolution (Q_f) in LID performances (EER(%)). Corresponding best performing time-domain WST EERs are denoted by dotted lines.*

4.3. Cross-corpora LID evaluation with optimized WST hyper-parameters

The cross-corpora performances for the three used corpora are presented in Table 2. As reported in different literature [24, 30, 31], we observe a prominent performance mismatch between the same-corpora and cross-corpora LID performances. However, we observe some improvements in cross-corpora LID performances with different WST hyper-parameters. Assume, for the train-test pair of corpus M and N ($M \neq N$), the best-reported cross-corpora LID performance is associated with WST hyper-parameter $(M\text{-}N)_{hp}$. Following this notation, in Table 2, we observe that $(M\text{-}N)_{hp} \neq (N\text{-}M)_{hp}$ for all three corpora. We also observe that $(M\text{-}N)_{hp} \neq N_{hp}$. Hence, the choice of optimal WST hyper-parameters depends on both train and test data. Hence, for deployment in unknown real-world scenarios, fusion of LID systems trained with WST extracted from different hyper-parameters are required.

4.4. Multi-WST LID system and blind evaluation

The optimal hyper-parameter sets for each corpus are decided based on their same-corpora LID performances. To eliminate

Table 3: *Blind evaluation (expressed as EER (%) / $C_{\mathrm{avg}} * 100$) on VoxLingua107 for the multi-WST fused LID systems.*

Training corpus	Evaluation corpus	Baseline MFCC	Best hp	Top-3 hp	Top-5 hp
IIITH	IIITH	9.74 / 11.51	7.53 / 8.20	**4.55 / 5.80**	4.89 / 6.10
	VoxLingua107	36.20 / 35.59	36.28 / 34.60	**32.88 / 32.07**	32.92 / **32.07**
LDC	LDC	21.86 / 25.70	12.87 / 15.41	8.65 / 10.20	**7.81 / 10.00**
	VoxLingua107	44.28 / 43.58	38.56 / 38.60	38.31 / 38.30	**37.88 / 37.70**
KGP	KGP	12.36 / 8.75	5.55 / 5.83	**4.87 / 4.20**	5.15 / 4.70
	VoxLinua107	42.72 / 42.56	**40.55** / 40.00	41.96 / **39.80**	41.36 / 40.20

any human-in-loop intervention in the final assessment, we conduct a blind evaluation approach using VoxLingua107 [21] corpus, following our earlier work [24]. We randomly select 500 utterances from each of the five languages. Repeating it four times, we create four blind test sets, each with 500 utterances. The average LID performances over all the blind test sets are reported in Table 3. For each training corpus, we consider the top-3 and top-5 WST hyper-parameter sets and fuse the corresponding LID systems. We use logistic regression based score fusion, [2] which are calibrated and trained with the validation set logits. The top-3 fusion systems prominently outperform the best-hp LID performance for IIITH and KGP. For LDC, top-5 fusion yields the best LID performance. From the MFCC baseline, the blind evaluation EER is decreased by 3.32%, 6.40%, and 2.17%, respectively, for the three training corpora.

5. Conclusions

To improve low-resourced LID generalization, we investigate wavelet scattering transform (WST) as an alternate feature representation. WST restores the high-frequency cues, which are lost in MFCCs with higher temporal context, as modulation spectrums. To our knowledge, this is the first work that explores WST for LID tasks. Experiments on multiple corpora show that LID tasks benefit the most with lower octave resolution in the first scattering layer. For news-read speech, smaller temporal context is desired, while the reverse holds true for conversational speech. We also find that frequency domain scattering is not beneficial for LID. Further, our cross-corpora experiments show that the optimal set of WST hyper-parameters is corpus-specific. Hence, we develop multi-WST fused LID systems for evaluation in unknown real-world scenarios. Compared to the MFCC-based baseline, the proposed system improves EER up to 14.05% and 6.40% for the same-corpora and blind cross-corpora evaluations, respectively. In future, we aim to develop an adaptive system to automatically obtain the optimal WST hyper-parameters depending on the data characteristics.

---

[2] https://gitlab.eurecom.fr/nautsch/pybosaris

6. References

[1] E. Ambikairajah *et al.*, "Language identification: A tutorial," *IEEE Circuits and Systems Magazine*, vol. 11, no. 2, pp. 82–108, 2011.

[2] S. Dey, M. Sahidullah, and G. Saha, "An overview on Indian spoken language recognition from machine learning perspective." *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 21, no. 128, pp. 1–45, 2022.

[3] S. Dey, G. Saha, and M. Sahidullah, "Cross-corpora language recognition: A preliminary investigation with Indian languages," in *EUSIPCO*. IEEE, 2021, pp. 546–550.

[4] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Prentice-Hall, Inc., 1993.

[5] J. Andén and S. Mallat, "Deep scattering spectrum," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4114–4128, 2014.

[6] H. Li, B. Ma, and K. A. Lee, "Spoken language recognition: from fundamentals to practice," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1136–1159, 2013.

[7] S. Mallat, "Group invariant scattering," *Communications on Pure and Applied Mathematics*, vol. 65, no. 10, pp. 1331–1398, 2012.

[8] C. Baugé, M. Lagrange, J. Andén, and S. Mallat, "Representing environmental sounds using the separable scattering transform," in *ICASSP*. IEEE, 2013, pp. 8667–8671.

[9] X. Y. Kek, C. S. Chin, and Y. Li, "An intelligent low-complexity computing interleaving wavelet scattering based mobile shuffling network for acoustic scene classification," *IEEE Access*, vol. 10, pp. 82 185–82 201, 2022.

[10] V. Hajihashemi, A. A. Gharahbagh, P. M. Cruz, M. C. Ferreira, J. J. Machado, and J. M. R. Tavares, "Binaural acoustic scene classification using wavelet scattering, parallel ensemble classifiers and nonlinear fusion," *Sensors*, vol. 22, no. 4, p. 1535, 2022.

[11] J. Bruna and S. Mallat, "Audio texture synthesis with scattering moments," *arXiv preprint arXiv:1311.0407*, 2013.

[12] N. M. Joy, D. Oglic, Z. Cvetkovic, P. Bell, and S. Renals, "Deep scattering power spectrum features for robust speech recognition." in *INTERSPEECH*. ISCA, 2020, pp. 1673–1677.

[13] V. Lostanlen, C. El-Hajj, M. Rossignol, G. Lafay, J. Andén, and M. Lagrange, "Time–frequency scattering accurately models auditory similarities between instrumental playing techniques," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2021, no. 1, pp. 1–21, 2021.

[14] C. Wang, E. Benetos, V. Lostanlen, and E. Chew, "Adaptive scattering transforms for playing technique recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1407–1421, 2022.

[15] P. Singh, G. Saha, and M. Sahidullah, "Deep scattering network for speech emotion recognition," in *EUSIPCO*. IEEE, 2021, pp. 131–135.

[16] W. Ghezaiel, L. Brun, and O. Lézoray, "Hybrid network for end-to-end text-independent speaker identification," in *International Conference on Pattern Recognition*. IEEE, 2021, pp. 2352–2359.

[17] Vuddagiri *et al.*, "IIITH-ILSC speech database for Indain language identification." in *Spoken Language Technologies for Under-Resourced Languages*. IEEE, 2018, pp. 56–60.

[18] J. Karen *et al.*, "Multi-language conversational telephone speech 2011 – South Asian LDC2017S14. web download. Philadelphia: Linguistic Data Consortium," -, 2017.

[19] S. Maity, V. A.K., K. Rao, and D. Nandi, "IITKGP-MLILSC speech database for language identification," in *National Conference on Communication (NCC)*. IEEE, 2012, pp. 1–5.

[20] D. Snyder *et al.*, "Spoken language recognition using x-vectors." in *Odyssey: The Speaker and Language Recognition Workshop*, 2018, pp. 105–111.

[21] J. Valk and T. Alumäe, "VoxLingua107: a dataset for spoken language recognition," in *Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 652–658.

[22] K. Mounika *et al.*, "An investigation of deep neural network architectures for language recognition in Indian languages." in *INTERSPEECH*. ISCA, 2016, pp. 2930–2933.

[23] T. Mandava, R. K. Vuddagiri, H. K. Vydana, and A. K. Vuppala, "An investigation of LSTM-CTC based joint acoustic model for Indian language identification," in *Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 389–396.

[24] S. Dey, M. Sahidullah, and G. Saha, "Cross-corpora spoken language identification with domain diversification and generalization," *Computer Speech & Language*, vol. 81, p. 101489, 2023.

[25] S. O. Sadjadi, T. Kheyrkhah, A. Tong, C. S. Greenberg, D. A. Reynolds, E. Singer, L. P. Mason, and J. Hernandez-Cordero, "The 2017 NIST language recognition evaluation." in *Odyssey: The Speaker and Language Recognition Workshop*. ISCA, 2018.

[26] Z. Li, M. Zhao, Q. Hong, L. Li, Z. Tang, D. Wang, L. Song, and C. Yang, "AP20-OLR challenge: Three tasks and their baselines," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2020, pp. 550–555.

[27] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *INTERSPEECH*. ISCA, 2020, pp. 1–5.

[28] P. Kumawat and A. Routray, "Applying TDNN architectures for analyzing duration dependencies on speech emotion recognition," in *INTERSPEECH*. ISCA, 2021, pp. 3410–3414.

[29] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.

[30] B. Schuller, B. Vlasenko, F. Eyben, M. Wöllmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-corpus acoustic emotion recognition: Variances and strategies," *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 119–131, 2010.

[31] B. Chettri, R. G. Hautamäki, M. Sahidullah, and T. Kinnunen, "Data quality as predictor of voice anti-spoofing generalization," in *INTERSPEECH*. ISCA, 2021, pp. 1659–1663.