IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 25, 2022, accepted February 5, 2022, date of publication February 9, 2022, date of current version February 17, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3150728

Differentiable Measures for Speech Spectral Modeling

MIGUEL ARJONA RAMÍREZ[1], (Senior Member, IEEE), WESLEY BECCARO[1], DEMÓSTENES ZEGARRA RODRÍGUEZ[2], (Senior Member, IEEE), AND RENATA LOPES ROSA[2]

[1]Department of Electronic Systems Engineering, Polytechnic School of the University of São Paulo, São Paulo 05508-010, Brazil
[2]Department of Computer Science, Federal University of Lavras, Lavras 37200-900, Brazil

Corresponding author: Miguel Arjona Ramírez (miguel@lps.usp.br)

This work was supported in part by the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) under Grant 2018/26455-8 and Grant 2019/07665-4.

ABSTRACT Autoregressive models for the envelope of speech power spectral densities (PSDs) are refined by the self-supervised spectral learning machine (S3LM) provided with differentiable spectral objective functions, including the Itakura-Saito divergence (ISD), the Kullback-Leibler divergence (KLD), the reverse KLD (RKLD) and the log spectral distortion (LSD), which display more significant results. However, in order to assess the models more perceptually, a method is proposed based upon perturbations around perfect reconstruction analysis-synthesis configurations. In the cross-excitation analysis-synthesis assessment (CEASA) method, the residual signals generated by analysis filters of the spectral models are injected as excitation into the synthesis filters derived from the same and other models in order to be evaluated by the perceptual evaluation of speech quality (PESQ) and Itakura divergence (ID), which are averaged over a set of models obtained using the objective functions mentioned above. The results lead to a superior performance when the RKLD is used as the loss function for the estimation of the spectral models with the ISD ranking close behind. The focus of these divergences on the spectral peaks is argued and pointed as the most important factor for this behavior. Specifically, using the PESQ scores obtained with CEASA, the RKLD loss is found to improve the performance by 1.0%, 4.0% and 19.3% with respect to the open-loop analysis, the KLD and the LSD models, respectively, while the corresponding improvements for the ISD loss are 0.1%, 3.0% and 18.2%, and the RKLD models excel the ISD models by 1.0% on average. Even though the spectral measures alone are not able to unequivocally distinguish the better of the two, CEASA is shown to have enough sensitivity to distinguish their performances. In summary, the learning machine S3LM fits models for the short-term spectral envelope of speech and, for the evaluation of its performance under several differentiable loss functions, the CEASA assessment tool has been developed. In addition, CEASA may be used for other assessments connected with speech analysis and synthesis.

INDEX TERMS Autoregressive processes, machine learning algorithms, prediction methods, self-supervised learning, speech analysis, spectral analysis.

I. INTRODUCTION

Models for the envelope of speech spectra [1] are important for various tasks that require speech analysis, such as speech coding, speech synthesis, automatic speech recognition and speech enhancement.

Autoregressive models for speech power spectral density $S(e^{j\omega})$ may be obtained by the application of the Wiener-Khinchin theorem to get the autocorrelation function [2]

$$R(m) = \frac{1}{2\pi}\int_{-\pi}^{\pi} S(e^{j\omega})e^{j\omega m}d\omega \tag{1}$$

for $m = 0, 1, \cdots, p$, in order to determine an autoregressive model of order p. This model may be obtained by the autocorrelation method of linear prediction, proposed by F. Itakura [3], [4]. The model may be represented by linear prediction coefficients [5] or by other transformed parameters. For instance, the analysis may require formant estimation and tracking [2], [6].

The associate editor coordinating the review of this manuscript and approving it for publication was Fu-Kwun Wang.

Despite the successful wide use of autoregressive analysis in speech applications [7], it has some shortcomings such as inaccuracies in modeling the discrete spectra arising in harmonic segments of speech [8], [9]. An interesting approach to harmonic spectral envelope estimation is true-envelope linear predictive coding (TE-LPC), which is an iterative cepstral technique based on a band-limited interpolation of the reference sub-sampled spectral envelope [10]. This work also proposes a residual spectral peak flatness measure for discrete spectra.

The shortcoming of straightforward autoregressive analysis of harmonic speech segments and other reasons motivate the improvement of autoregressive spectral estimation by means of machine learning methods. For instance, Cui *et al.* [9] show that adaptive changes performed by a deep neural network (DNN) to the spectra to be analyzed improve the quality of supervised spectral models.

Also, models for speech spectral envelopes play a significant role in speech synthesis, where a major problem is the oversmoothing of the reconstructed spectral envelopes [11]. In order to ameliorate this effect, restricted Boltzmann machines and deep belief networks have been proposed for modeling spectral envelopes [12]. It is also important to note that spectral envelope features can be efficiently detected by means of unsupervised methods [6].

Spectral envelopes may alternatively be obtained by means of cepstral coefficients as in this application of machine learning to speech emotion recognition [13]. In addition, mel frequency cepstral coefficients (MFCCs) are also reported to be used in emotional speech synthesis [14].

In the performance evaluation of diverse speech solutions or applications [15], speech quality assessment methods are widely adopted. For instance, in [16], a complex spectral mapping based on DNN is proposed, and its results were evaluated using the algorithm described in ITU-T Rec. P.862 [17], [18], mostly known as PESQ. Another speech quality metric is the Virtual Speech Quality Objective Listener, known as ViSQOL [19], that uses spectral and temporal parameters to determine a listening quality objective (LQO) score using the 5-point quality scale. In connection with these applications, we propose an analysis-synthesis assessment method for the spectral models which is more suitable to evaluate their performance in action.

In this context, this work intends to improve the open-loop analytical (OLA) model using a machine learning algorithm in conjunction with several differentiable loss functions that are applied to the reference and reconstructed power spectral densities. The differentiable losses implemented in the S3LM architecture and used in the experimental tests were the ISD [3], [20], the KLD, the reverse KLD, and the LSD. For each loss function, S3LM produces a distinctive spectral envelope model. The cross-excitation analysis-synthesis assessment (CEASA) was used to jointly assess the fidelity of the spectral envelope models considered in the tests by means of the synthesized speech signals obtained from the parameters associated to each model. In summary, each spectral model is used as two filters, namely, an analysis filter and a synthesis filter, which are associated in cascade. Further, the input to the synthesis filter is alternatively provided by the output of the analysis filter of the corresponding model and also by each of the other models whereas the reference signal is input to the analysis filter. Finally, in order to perform a better quality analysis, the synthesized signal is compared with the reference signal using both the PESQ and the ID [20], [21] algorithms. This procedure is carried out for all combinations of analysis and synthesis filters for all spectral model pairs generated with different losses for the spectrum of the same reference signal. In addition, two different window sizes for the speech signal are used to obtain spectral models that, beyond allowing one to analyze the impact of window length on the spectral fitting measures for the spectral models, also underlines the need for a nonspectral assessment tool such as CEASA. This independent assessment is necessary because CEASA tends to amplify the distinction between different models and also dismisses seeming static spectral fit improvements brought about by window length change, which turn out to be illusory.

Nowadays, different solutions based on both signal processing methods and machine learning algorithms are applied in several research areas [22], [23]. In the present work, we use signal processing techniques such as autoregressive models, prediction and perfect reconstruction in analysis-synthesis systems which are integrated with machine learning structures to come up with tied spectral weighting layers (TSWLs). These techniques are used both in the proposed learning machine for the layers and the losses and in the CEASA diagnostic tool which includes analysis-synthesis techniques based on perfect reconstruction.

It is noted that the CEASA assessment tool is intended to be used with rather high quality spectral models since it should cause its analysis-synthesis system to operate around the perfect reconstruction condition. Further, under these conditions, PESQ-LQO is a trustworthy quality score, whose results are also corroborated with those given by the VisQOL metric.

Addressing the issues raised above, this article presents the proposed S3LM in Section II, the most important measures for speech spectral analysis in Section III, the spectral measures used as loss functions and the comparison of the spectral models they lead to in Section IV and the description of the CEASA method along with the results of its application to the speech spectral models in Section V. Finally, the major results in this article are connected in conclusion in Section VI.

II. THE SELF-SUPERVISED SPECTRAL LEARNING MACHINE

As previously stated, we propose a learning machine that inputs a spectrogram as a sequence of one-sided log PSDs with K samples up to the Nyquist frequency for an $F_s = 16$ kHz sampling rate.

The network architecture of the proposed S3LM is composed by three tied spectral weighting layers (TSWLs), as shown in Fig. 1, with tied weight vector $\boldsymbol{w}_0$ and tied bias

vector $\boldsymbol{b}_0$, both the size K of the PSDs, which are extended over the spectrogram for each training epoch.

The S3LM architecture performs spectral pre-processing using the TSWLs. The structure consists of artificial neurons applied to each spectral component. Rather than fully connected networks, the proposed model has a singly connected architecture with two hidden layers and the weights shared between the layers. This structure concentrates attention on each spectral bin for closer convergence up to the same number of epochs while, at the same time, the strategy also brings about a reduction in the number of parameters and training time.

We will represent a single log PSD as

$$P_L(k) = 10\log_{10} S\left(\exp\left(j2\pi\frac{k}{2(K-1)}\right)\right) \tag{2}$$

for $k = 0, 1, \cdots, K-1$, which forward propagates through the first three layers as

$$\begin{aligned}\boldsymbol{h}_0 &= \phi\left(\boldsymbol{w}_0 \circ \boldsymbol{P}_L + \boldsymbol{b}_0\right)\\ \boldsymbol{h}_1 &= \phi\left(\boldsymbol{w}_0 \circ \boldsymbol{h}_0 + \boldsymbol{b}_0\right)\\ \boldsymbol{h}_2 &= \phi\left(\boldsymbol{w}_0 \circ \boldsymbol{h}_1 + \boldsymbol{b}_0\right),\end{aligned} \tag{3}$$

where $\phi(\cdot)$ is the rectified linear unit (ReLU) activation function, $\circ$ represents the Hadamard or elementwise product, and $\boldsymbol{h}_0$, $\boldsymbol{h}_1$, and $\boldsymbol{h}_2$ are the outputs of each weighting layer.

So the modified log PSD is $\boldsymbol{h}_2$, resulting in the modified PSD obtained as

$$P_2(k) = 10^{h_2(k)/10} \tag{4}$$

for $k = 0, 1, \cdots, K-1$. And now the modified autocorrelation function is obtained by using the Wiener-Khinchin theorem [24] as

$$\begin{aligned}R(m) = P_2(0) + 2\sum_{k=1}^{K-2} P_2(k)\cos\left(\frac{2\pi}{2(K-1)}km\right)\\ + P_2(K-1)\cos(\pi m)\end{aligned} \tag{5}$$

for $m = 0, 1, \ldots, p$. From these autocorrelation coefficients a prediction analysis is performed, leading to the prediction coefficient vector

$$\boldsymbol{a} = \begin{bmatrix}1 & a_1 & \cdots & a_p\end{bmatrix}^T. \tag{6}$$

For a general prediction coefficient vector $\boldsymbol{\alpha}$, the square prediction error is

$$\epsilon_{\boldsymbol{\alpha}} = \boldsymbol{\alpha}^T\boldsymbol{R}\boldsymbol{\alpha}, \tag{7}$$

where $\boldsymbol{R}$ is the $(p+1)\times(p+1)$ Toeplitz reference augmented autocorrelation matrix whose entries are given by (5). For the special prediction coefficient vector $\boldsymbol{\alpha} = \boldsymbol{a}$, the minimum prediction error achieved is

$$\begin{aligned}\epsilon_{\min} &= \boldsymbol{a}^T\boldsymbol{R}\boldsymbol{a}\\ &= \boldsymbol{a}^T\boldsymbol{r},\end{aligned} \tag{8}$$

where $\boldsymbol{r} = \begin{bmatrix}R(0) & R(1) & \cdots & R(p)\end{bmatrix}^T$.

After linear prediction analysis, the reconstructed PSD [4] is obtained as

$$\widetilde{P}_2(k) = \frac{\epsilon_{\min}}{\left|1 + \sum_{\ell=1}^{p} a_\ell \exp\left(-j\frac{2\pi}{2(K-1)}k\ell\right)\right|^2} \tag{9}$$

for $k = 0, 1, \cdots, K-1$ or, alternatively, by

$$\widetilde{P}_2(k) = \frac{\epsilon_{\min}}{\sum_{m=-p}^{p} R_{aa}(m)\exp\left(j\frac{2\pi}{2(K-1)}km\right)} \tag{10}$$

for $k = 0, 1, \cdots, K-1$, where the autocorrelation function of the linear prediction vector is

$$R_{aa}(m) = \sum_{\ell=-p}^{p} a_\ell a_{\ell+m}, \tag{11}$$

where $a_\ell = 0$ for $\ell < 0$ or $\ell > p$ and $a_o = 1$. Equation (9) is arguably simpler than Eq. (10) for gradient backpropagation.

Then, the log reconstructed PSD is obtained as

$$\widetilde{h}_2(k) = 10\log_{10}\widetilde{P}_2(k) \tag{12}$$

for $k = 0, 1, \cdots, K-1$ and either the PSD or the log PSD, $\widetilde{\boldsymbol{h}}_2$, may be used for computing the loss function according to the arguments of this function.

The model is implemented using the deep learning framework PyTorch. The weights $\boldsymbol{w}_0$ of S3LM are initialized to all ones while its biases $\boldsymbol{b}_0$ are all initialized from samples of a zero-mean Gaussian distribution with a standard deviation $\sigma = 1\times10^{-4}$ and they are optimized by a stochastic gradient descent algorithm with a learning rate $\ell_r = 1\times10^{-4}$. Good convergence has been observed after 80 epochs.

The model was experimented using the TIMIT Acoustic-Phonetic Continuous Speech Corpus dataset [25]. TIMIT has 6300 utterances (10 sentences spoken by each of 630 speakers) separated in 16 bit-wav files with a sampling rate of 16 kHz. The speakers are distributed across 8 different dialect regions. The utterances in dialect regions 1 through 4 of the test set of TIMIT dataset were selected for modeling K-sample one-sided log PSDs by self-supervised methods with $K = 1025$.

We used a workstation computer with 16 GB of RAM, an Intel® Xeon® E-2146G CPU at 3.50 GHz with 6 cores, and a single NVIDIA GPU card with 4 GB. Training and testing are simultaneous since S3LM is self-supervised.

III. MEASURES FOR SPECTRAL ANALYSIS

Based on square prediction errors, an important measure for comparing autoregressive models is Itakura divergence (ID). For the reference autoregressive vector $\boldsymbol{a}$ and the estimated vector $\widetilde{\boldsymbol{a}}$, Itakura divergence [20], [21] is given by

$$\begin{aligned}D_{\mathrm{I}}(\boldsymbol{a},\widetilde{\boldsymbol{a}}) &= \frac{\epsilon_{\widetilde{\boldsymbol{a}}}}{\epsilon_{\min}}\\ &= \frac{\widetilde{\boldsymbol{a}}^T\boldsymbol{R}\widetilde{\boldsymbol{a}}}{\boldsymbol{a}^T\boldsymbol{R}\boldsymbol{a}}.\end{aligned} \tag{13}$$

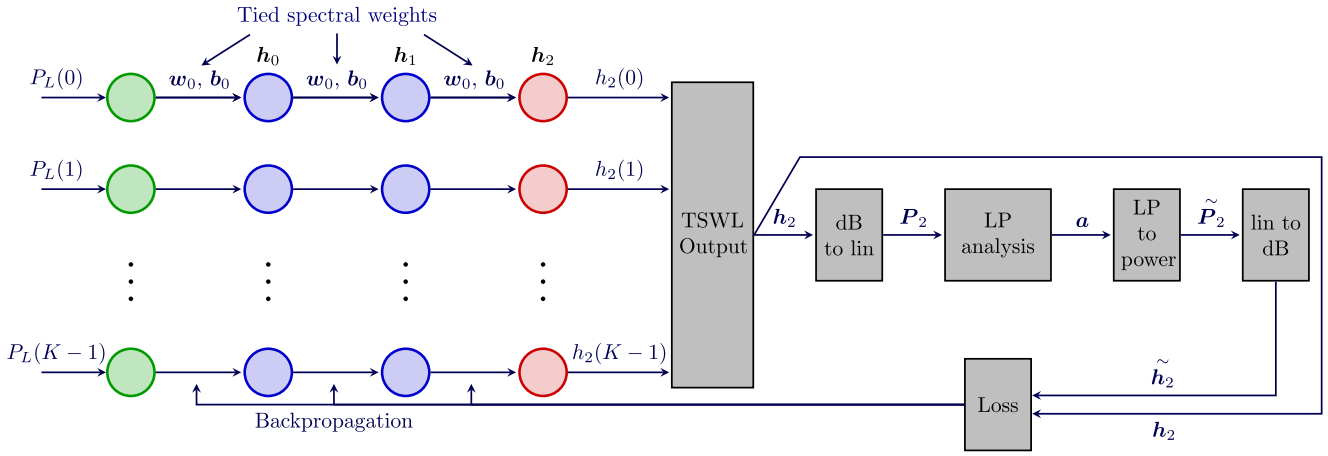

FIGURE 1. Architecture of the proposed self-supervised spectral learning machine (S3LM).

This definition, originally called "likelihood ratio" by Itakura [21], makes it clear that the minimum possible value for ID is unity, corresponding to the minimum square prediction error condition, therefore coinciding with the result for open-loop linear prediction analysis. On the other hand, it does not have any inherent upper bound, even though a practical value of 1.4 has been mentioned as the frontier beyond which synthetic speech quality is too low to be useful [26].

However, in order to be used as a loss function in comparing PSDs, the Itakura-Saito divergence (ISD) is more straightforward than ID and it is defined by [3], [20]

$$D_{\mathrm{IS}}(P, Q) = \frac{1}{f_{\mathrm{Ny}}} \int_0^{f_{\mathrm{Ny}}} \left[\frac{P(f)}{Q(f)} - \log \frac{P(f)}{Q(f)} - 1 \right] df, \tag{14}$$

where $P(f)$ is the reference PSD, $Q(f)$ is the distorted or reconstructed PSD and f_{Ny} is Nyquist frequency. For sampled PSDs, the ISD is given by

$$D_{\mathrm{IS}}(P, Q) = \frac{1}{K} \sum_{k=0}^{K-1} \left[\frac{P(k)}{Q(k)} - \log \frac{P(k)}{Q(k)} - 1 \right]. \tag{15}$$

In Section II, where the proposed S3LM was described, we have reference PSD as $\boldsymbol{P}_2$ and reconstructed PSD as $\widetilde{\boldsymbol{P}}_2$.

A more general spectral distortion measure which is not conceived for measuring autoregressive spectral fit in particular is the log-spectral distortion (LSD), which is expressed in dB as

$$D_{\mathrm{LS}}(P, Q) = \sqrt{\frac{1}{f_{\mathrm{Ny}}} \int_0^{f_{\mathrm{Ny}}} \left[10 \log_{10} \frac{P(f)}{Q(f)} \right]^2 df}. \tag{16}$$

Notwithstanding their different constitutions, it is interesting to observe that both the square error and the ISD are instances of Bregman divergences [27], which also holds as a class member the generalized Kullback-Leibler divergence (GKLD), defined by

$$D_{\mathrm{GKL}}(p \,||\, q) = \frac{1}{K} \left(\sum_{k=0}^{K-1} p(k) \log \frac{p(k)}{q(k)} - \sum_{k=0}^{K-1} p(k) + \sum_{k=0}^{K-1} q(k) \right). \tag{17}$$

In Machine Learning, it is usual to employ probability density functions (PDFs) or probability mass functions (PMFs). First, we observe that PSDs are nonnegative and, while log PSDs may take on negative values, they may be raised to 0 dB by subtracting the minimum value from the whole log-spectrum. Second, if we normalize the log PSDs so that they sum to unity, then the GKLD reduces to the KLD, the Kullback-Leibler divergence. The possibility of processing PSDs just as PDFs for KLD measures and modeling has already been pointed out by [28]. The KLD from PDF q to PDF p is defined as

$$D_{\mathrm{KL}}(p \,||\, q) = \int_{\mathbb{R}^D} p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} d\boldsymbol{x}, \tag{18}$$

as long as $S_p \subseteq S_q$, where S_p and S_q are, respectively, the supports for the D-variate PDFs p e q, avoiding the occurrence of infinities [14] for points where $q(x) = 0$ and $p(x) > 0$ in (18). For PMFs, the KLD from q to p is computed as

$$D_{\mathrm{KL}}(p \,||\, q) = \sum_{k=0}^{K-1} p(k) \log \frac{p(k)}{q(k)}, \tag{19}$$

which represents the direct KLD as long as p is a data PMF and q is a latent variable PMF. By keeping their roles while exchanging the positions of p and q in the argument of the divergence function, we obtain the reverse KLD (RKLD) as

$$D_{\mathrm{KL}}(q \,||\, p) = \sum_{k=0}^{K-1} q(k) \log \frac{q(k)}{p(k)}. \tag{20}$$

IV. DIFFERENTIABLE LOSS COMPARISONS

The open loop analytical (OLA) analysis based on the autocorrelation method is used as a baseline for assessing the refinements brought about by the learning methods. Its objective function is a square prediction error, which is a square distance in polynomial space provided with a time-varying inner product defined by the short-term autocorrelation function [8].

The differentiable losses that have been applied to the reference and reconstructed power spectral densities (PSDs) are the Itakura-Saito divergence (ISD), the Kullback-Leibler divergence (KLD), the reverse KLD (RKLD) and the log spectral distortion (LSD). Most of these measures are also used for assessing the reconstructed PSDs, including in addition Jeffrey's divergence (JD) [20], which provides a balance between the KLD and the RKLD as a measurement tool.

It is interesting to observe that it can be demonstrated that the minimization of the ISD with respect to the prediction coefficients is equivalent to the minimization of the square prediction error in polynomial space [4]. However, it may be argued that the path leading to the minimum may be different in an iterative approach.

Differentiable signal processing methods have made it possible to perform the short-time Fourier transform (STFT) with variable hop size windows [29]. This research has lead us to discover some interesting spectral fitting differences that depend on STFT window length.

All our PSDs have been obtained using sequences of 50% overlapping sine windows.

The performance of the various methods for female speakers and 20 ms long windows are shown in Table 1, where quality improvements (QI) are positive for a result greater than the OLA baseline when it is a quality or similarity measure and are also positive for a result smaller than the OLA baseline when it is about a divergence or distortion measure. More precisely, the quality improvement for measure M is computed as

$$\mathrm{QI}(M) = \pm \left(M\left(P_{\mathrm{M}}, P_{\mathrm{ref}}\right) - M\left(P_{\mathrm{OLA}}, P_{\mathrm{ref}}\right)\right), \qquad (21)$$

where the plus sign is selected if M is a quality measure while the minus sign is selected if M is a divergence measure, P_{M} is the PSD for the model obtained with M as objective function, P_{OLA} is the PSD for the model obtained by the open-loop analysis and P_{ref} is the reference power spectral density.

The utterances in dialect regions 1 through 4 of the test set of TIMIT speech corpus [25] were selected for modeling K-sample one-sided log PSDs by self-supervised methods with $K = 1025$.

Using the self-supervised learning machine several measures are used as objective function alternatively as shown in Table 1 for female speakers and long windows in its leftmost column and the measures appearing as headers for the next columns are alternative measures for the comparisons between the PSDs obtained and the corresponding reference PSDs. Also, the same is shown for male talkers and long windows in Table 2, for female speakers and short windows in Table 3 and for male speakers and short windows in Table 4. However, short windows have been used only in a preliminary way for a couple of utterances.

By observing the results in the abovementioned tables it stands out that ISD is the only objective function that can make the learning machine improve the quality under the ISD measure, which is arguably the most significant measure for speech PSD envelopes. The ISD objective function also brings about quality improvement that can be seen by the LSD measure. On the other hand, the KLD objective function, which is widely used in Machine Learning, can consistently improve quality as measured by both the KLD and the JD and even, in most cases, its quality improvement is also seen by the LSD, particularly in Tables 1 and 2, but fails in Table 4.

The RKLD objective function may cause quality improvements to be detected by the KLD, the JD and the LSD measures in Table 1, even exceeding the quality improvement of the KLD as seen by itself in Tables 1 and 2, but it may also fail to have any quality improvement seen by any of those three measures as happens in Table 3. A similar behavior is displayed by the LSD objective function, which is able to cause quality improvements detectable by the KLD, the JD and the LSD measures in Table 1 and Table 2 but quality improvements fail to be seen by the KLD and the JD in Table 3. But a final analysis about these apparent shortcomings of the RKLD should be postponed till a more complete performance assessment is disclosed in Section V.

Further, the good performance of the RKLD is only partially offset by its not so good performance as measured by the ISD even if it is still the best performing loss as seen by the ISD in the set of losses that includes also the KLD and the LSD.

Finally, the LSD models behave in a rather contradictory manner, being the worst as measured by the ISD measure but beating the models obtained with the other losses by the greatest margin in several instances.

As a final overall observation, absolute scores are seen to improve for short spectral estimation windows when compared with long windows and particularly so when measured by the ISD measure. The improvement is also significant when measured by the LSD except when the same LSD is used as a loss function as well for male speaker. In this case, when the LSD is used as the loss function, the KLD and the JD also fail to notice any improvement.

V. ASSESSMENT RESULTS

In order to assess the fidelity of the spectral envelope model in more neutral conditions, the cross-excitation analysis-synthesis assessment (CEASA) was used, which is depicted in Fig. 2 for the simple case involving two models, where two prediction vectors $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ are input from S3LM or any other modeling system for that matter. In its turn, CEASA uses the prediction vectors to come up with the analysis filters

$$A_1(z) = \sum_{i=0}^{p} a_{1i} z^{-i} \qquad (22)$$

TABLE 1. Average measures for speech obtained from spectral models generated by methods based on five loss functions for female speakers and quality improvements (QI) over OLA. Spectra are obtained for 20 ms long windows.

Method (Loss)	ISD	QI(ISD)	KLD	QI(KLD)	JD	QI(JD)	LSD (dB)	QI(LSD) (dB)
OLA	0.7579	0	0.2301	0	0.3625	0	7.5179	0
ISD	0.7579	5.56×10^{-6}	0.2350	-4.92×10^{-3}	0.3674	-4.89×10^{-3}	7.5169	1.08×10^{-3}
RKLD	0.7604	-2.52×10^{-3}	0.2300	52.80×10^{-6}	0.3588	3.74×10^{-3}	7.4646	53.36×10^{-3}
KLD	0.8451	-87.23×10^{-3}	0.2234	6.67×10^{-3}	0.3485	14.02×10^{-3}	7.3102	207.70×10^{-3}
LSD	1.0887	-330.80×10^{-3}	0.1859	44.23×10^{-3}	0.2878	74.76×10^{-3}	5.6225	1.8954

TABLE 2. Average measures for speech obtained from spectral models generated by methods based on five loss functions for male speakers and quality improvements (QI) over OLA. Spectra are obtained for 20 ms long windows.

Method (Loss)	ISD	QI(ISD)	KLD	QI(KLD)	JD	QI(JD)	LSD (dB)	QI(LSD) (dB)
OLA	0.6113	0	0.2250	0	0.3523	0	6.5607	0
ISD	0.6113	7.73×10^{-7}	0.2298	-4.87×10^{-3}	0.3572	-4.85×10^{-3}	6.5601	662.90×10^{-6}
RKLD	0.6132	-1.87×10^{-3}	0.2246	403.63×10^{-6}	0.3487	3.64×10^{-3}	6.5367	24.07×10^{-3}
KLD	0.6706	-59.28×10^{-3}	0.2187	6.30×10^{-3}	0.3398	12.47×10^{-3}	6.4083	152.46×10^{-3}
LSD	0.6608	-49.49×10^{-3}	0.1733	51.65×10^{-3}	0.2686	83.74×10^{-3}	4.7091	1.8517

TABLE 3. Average measures for speech obtained from spectral models generated by methods based on five loss functions for female speakers and quality improvements (QI) over OLA. Spectra are obtained for 7.25 ms long windows.

Method (Loss)	ISD	QI(ISD)	KLD	QI(KLD)	JD	QI(JD)	LSD (dB)	QI(LSD) (dB)
OLA	0.4260	0	0.1956	0	0.2985	0	5.3403	0
ISD	0.4260	2.86×10^{-6}	0.2004	-4.8×10^{-3}	0.3033	-4.8×10^{-3}	5.3392	1.1×10^{-3}
RKLD	0.4369	-10.9×10^{-3}	0.2075	-11.9×10^{-3}	0.3175	-19.0×10^{-3}	5.3539	-13.6×10^{-3}
KLD	0.4447	-18.8×10^{-3}	0.1951	500.0×10^{-6}	0.2932	5.3×10^{-3}	5.3156	24.7×10^{-3}
LSD	0.5239	-98.0×10^{-3}	0.1984	-2.8×10^{-3}	0.2987	-200.0×10^{-6}	5.0360	304.3×10^{-3}

TABLE 4. Average measures for speech obtained from spectral models generated by methods based on five loss functions for male speakers and quality improvements (QI) over OLA. Spectra are obtained for 7.25 ms long windows.

Method (Loss)	ISD	QI(ISD)	KLD	QI(KLD)	JD	QI(JD)	LSD (dB)	QI(LSD) (dB)
OLA	0.4326	0	0.2072	0	0.3201	0	5.3685	0
ISD	0.4326	1.4305×10^{-6}	0.2119	-4.7×10^{-3}	0.3248	-4.7×10^{-3}	5.3677	800.0×10^{-6}
RKLD	0.4369	-4.3×10^{-3}	0.2075	-300.0×10^{-6}	0.3175	2.6×10^{-3}	5.3539	14.6×10^{-3}
KLD	0.4655	-32.9×10^{-3}	0.2000	7.2×10^{-3}	0.3058	14.3×10^{-3}	5.3952	-26.7×10^{-3}
LSD	0.6179	-185.3×10^{-3}	0.2100	2.8×10^{-3}	0.3207	600.0×10^{-6}	5.0718	296.7×10^{-3}

and

$$A_2(z) = \sum_{i=0}^{p} a_{2i} z^{-i}, \tag{23}$$

where p is the order of the models and the synthesis filters are obtained as $H_1(z) = 1/A_1(z)$ and $H_2(z) = 1/A_2(z)$.

For a given speech signal $s(n)$ and a spectral model, the speech signal is injected into the corresponding analysis filter whose output is its residual signal, either $e_1(n)$ or $e_2(n)$, which is injected into both synthesis filters. As a result, different synthesized signals are obtained that are represented by $s_{11}(n)$, $s_{12}(n)$, $s_{21}(n)$, and $s_{22}(n)$. These synthesized signals, which provide a realization of their corresponding spectral models, are assessed by the PESQ algorithm, which provides a mean opinion score listening quality objective (MOS-LQO) measure [17], [18], and the Itakura divergence (ID) [20], [21]. Both measures are represented by the block named Meas(·) depicted in Fig. 2. By using each spectral model in turn for the analysis filter, two sets of measures are obtained for each synthesis filter and the mean value of each set of values is ascribed to the spectral model of the corresponding synthesis filter.

The basis for the operation of CEASA analysis-synthesis filter cascade is the perfect reconstruction condition which prevails when both the analysis filter and the synthesis filter in the cascade connection are configured with the same prediction vector so that the synthesized signal will coincide with the input signal up to a time delay in the absence of numerical errors.

After investigating the application of different window lengths in spectral modeling, the divergences were found to decrease for shorter windows as reported in Section IV. So we have decided to test the modeling algorithms for longer 20 ms

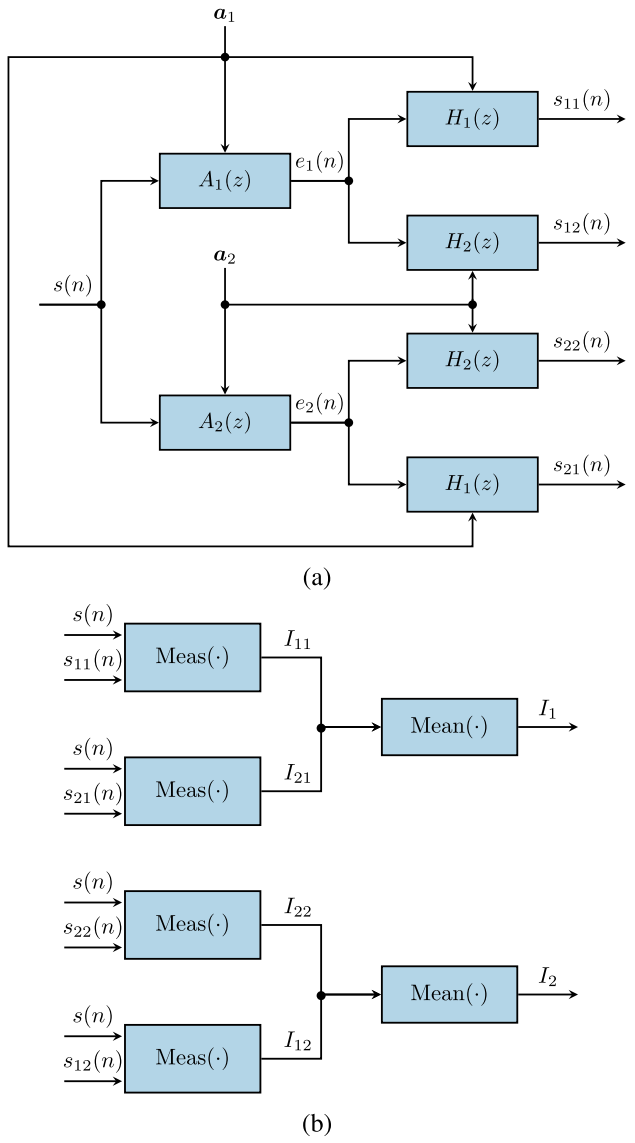

FIGURE 2. Cross-excitation analysis-synthesis assessment illustrated for two analysis-synthesis models: (a) Analysis cascaded with self-synthesis and cross-synthesis. (b) Association of performance scores, I, for measure Meas(·), and mean calculation for each synthesis filter.

windows over the dialect regions 1 through 4 of the test set of the TIMIT corpus [25] for female and male speakers while shorter 7.25 ms windows have been tested only for a couple of speakers due to their CEASA scores to be reported below.

In Table 5, longer windows are used for the spectral modeling of the utterances of female speakers, where ISD displays a small, however consistent, better performance which can be checked for the case of shorter windows from female utterances in Table 8 as well as male utterances in Tables 6 and 9 for longer and shorter windowing.

In order to check the confidence of the results, we have also assessed them using the ViSQOL measure [19], whose scores and attendant quality improvements for the machine learning methods over OLA are reported in Table 7, which, upon further comparison between methods and ranking of methods, is consistent with Tables 5 and 6.

However, if we keep to longer windows, the best performing loss is the RKLD, either assessed by PESQ or ID. This best performance within this set of losses is hinted by a qualitative analysis of the defining equation of the RKLD (20) in comparison to the defining equation of the KLD (19). As the weighting coefficients for the RKLD are the reconstructed masses $q(k)$, when RKLD is used as the loss, $q(k)$ should converge to small values in regions where the data masses $p(k)$ are rather small and, by themselves, would increase the argument of the log function unless $q(k)$ converges to a comparable small value. This would lead $q(k)$ to place most of its probability mass near the peaks of $p(k)$ instead, which is a good behavior to be valued by the ISD measure. An illustrated discussion of this convergence behavior under the minimization of the RKLD may be found in [30], where the equations for the divergences are the same as the abovementioned ones but the labels RKLD and KLD are exchanged.

Besides, as a matter of fact, shorter windows are found by CEASA to lead to lower performance than longer windows, contrary to what happens for pure spectral analysis in Section IV. This behavior is due to the dynamics of the synthesis filter in the assessment procedure. Further, it seems to indicate that shorter windows may be better for some spectral analysis tasks but longer windows are recommended for synthesis and related tasks that end up generating a speech signal.

As a curious outcome, we may find it surprising that the LSD models, which performed very well for all the measures but for the ISD, have ranked last in both the PESQ and ID scores for long windows. This highlights the fact that spectral envelope models should be better in matching spectral peaks than overall spectral details and this is captured more clearly by CEASA assessment than by static spectral measures.

It is noticeable by comparing the scores in Tables 5 and 6 that the spectral envelope models for male speakers fit their references more closely than those for female speakers. This behavior can be further checked by contrasting the box plots in Fig. 3. While modeling the spectral envelopes should not be affected by the local harmonic structure of the spectrum, this is valid when the density of harmonics is high enough so that the spectrum is approximately continuous. The latter is the condition for a lower pitched speaker, which is usually the case of a male speaker, which is consistent with the observation. Nonetheless, by referring again to the two tables mentioned above, we notice that the performances of the loss functions are ranked in the same order, irrespective of whether the speakers are female or male.

In short, using the PESQ scores obtained with CEASA, the RKLD loss is found to improve the performance by 1.0%, 4.0% and 19.3% with respect to the open-loop analysis, the KLD and the LSD models, respectively, while the corresponding improvements for the ISD loss are 0.1%, 3.0% and 18.2% and the RKLD models excel the ISD models by 1.0% on average. In a different approach to spectral envelope

TABLE 5. Average PESQ MOS-LQO quality and Itakura distortion from CEASA measures for speech obtained from spectral models generated by methods based on five loss functions for female speakers. Quality improvements over OLA are also presented. Spectra are obtained for 20 ms long windows.

Method	PESQ (MOS-LQO)	QI(PESQ)	Itakura divergence (ID)	Q(ID)
OLA	3.8916	0	1.0180	0
ISD	3.8949	3.23×10^{-3}	1.0179	64.06×10^{-6}
RKLD	3.9036	11.98×10^{-3}	1.0136	4.42×10^{-3}
KLD	3.7584	-133.25×10^{-3}	1.0483	-30.30×10^{-3}
LSD	3.1241	-767.54×10^{-3}	1.0675	-49.51×10^{-3}

TABLE 6. Average PESQ MOS-LQO quality and Itakura distortion from CEASA measures for speech obtained from spectral models generated by methods based on five loss functions for male speakers. Quality improvements over OLA are also presented. Spectra are obtained for 20 ms long windows.

Method	PESQ (MOS-LQO)	QI(PESQ)	Itakura divergence (ID)	QI(ID)
OLA	4.2370	0	1.0070	0
ISD	4.2382	1.25×10^{-3}	1.0070	25.56×10^{-6}
RKLD	4.3110	73.95×10^{-3}	1.0046	2.40×10^{-3}
KLD	4.1429	-94.08×10^{-3}	1.0124	-5.37×10^{-3}
LSD	3.7946	-442.37×10^{-3}	1.0185	-11.46×10^{-3}

TABLE 7. Average ViSQOL MOS-LQO scores from CEASA setup for speech obtained from spectral models generated by methods based on five loss functions for female and male speakers. Quality improvements over OLA are also presented. Spectra are obtained for 20 ms long windows.

Method	ViSQOL female (MOS-LQO)	QI(ViSQOL)	ViSQOL male (ID)	QI(ViSQOL)
OLA	4.8276	0	4.9138	0
ISD	4.8285	838.85×10^{-6}	4.9145	681.91×10^{-6}
RKLD	4.8809	53.23×10^{-3}	4.9512	37.43×10^{-3}
KLD	4.8244	-3.23×10^{-3}	4.9133	-495.35×10^{-6}
LSD	4.6615	-166.15×10^{-3}	4.8797	-34.09×10^{-3}

TABLE 8. Average PESQ MOS-LQO quality and Itakura distortion from CEASA measures for speech obtained from spectral models generated by methods based on five loss functions for female speakers. Quality improvements over OLA are also presented. Spectra are obtained for 7.25 ms long windows.

Method	PESQ (MOS-LQO)	QI(PESQ)	Itakura divergence (ID)	QI(ID)
OLA	3.3374	0	1.0260	0
ISD	3.3466	9.20×10^{-3}	1.0254	600.00×10^{-6}
RKLD	2.4792	-858.20×10^{-3}	1.0600	-34.00×10^{-3}
KLD	3.0610	-276.40×10^{-3}	1.0167	9.30×10^{-3}
LSD	3.2270	-110.40×10^{-3}	1.0224	3.60×10^{-3}

TABLE 9. Average PESQ MOS-LQO quality and Itakura distortion from CEASA measures for speech obtained from spectral models generated by methods based on five loss functions for male speakers. Quality improvements over OLA are also presented. Spectra are obtained for 7.25 ms long windows.

Method	PESQ (MOS-LQO)	QI(PESQ)	Itakura divergence (ID)	QI(ID)
OLA	3.8150	0	1.0180	0
ISD	3.8464	31.40×10^{-3}	1.0177	300.00×10^{-6}
RKLD	3.7438	-71.20×10^{-3}	1.0138	4.20×10^{-3}
KLD	3.2006	-614.40×10^{-3}	1.0338	-15.80×10^{-3}
LSD	3.4004	-414.60×10^{-3}	1.0294	-11.40×10^{-3}

modeling [9], three different supervised deep neural networks have been proposed, namely, DNNs, DNN and DNN1. The three deep neural networks have been trained with several divergences as loss functions and their models have been evaluated using the same function as the metric. Their best result is found for the DNN1 with the KLD, where it excels OLA by 93% measured by the KLD and the improvements over DNNs and DNN are 40% and 13%, respectively. On the other hand, when the ISD is used, DNN falls behind OLA by 2% and is also worse than DNNs and DNN by 2% and 1%, respectively.

As can be observed in the experimental results and their analysis, the methodology proposed in this work achieves a significant improvement in the sensitivity of the assessment of fitting for different spectral envelope models.

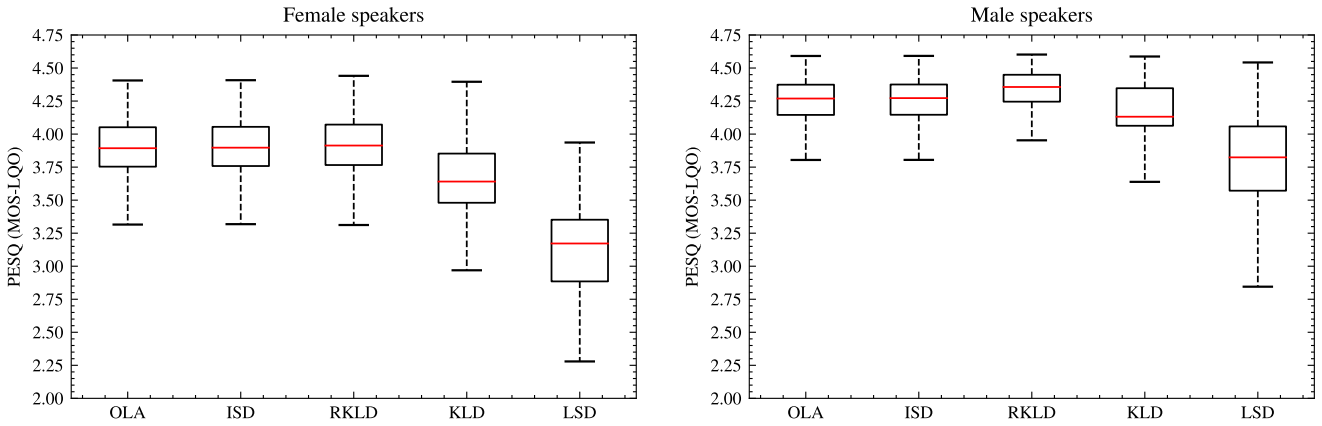

FIGURE 3. Box plots for the distributions of PESQ scores corresponding to Tables 5 (female speakers) and 6 (male speakers) with boxes extending from the first quartile below to the third quartile above, including the median line in red and two whiskers reaching out to the minimum and maximum values.

VI. CONCLUSION

Spectral envelope models for speech signals have relied for quite some time on linear prediction analysis. This work proposes a refinement to open-loop analytical (OLA) models by using machine learning algorithms provided with differentiable losses. Losses that have been proposed previously in autoregressive analysis are investigated for this task as well as popular divergences used in machine learning. Since the results obtained by spectral measures are not conclusive at first as to the most suitable losses, a quality assessment method is proposed based on the fundamental perfect reconstruction criterion for cascaded analysis-synthesis systems. Using the original speech signals and the analysis and synthesis filters defined by the parameters of the spectral envelope models, all possible analysis-synthesis cascades are mounted in the proposed cross-excitation analysis-synthesis assessment (CEASA) method. For the whole set of signals, the reverse Kullback-Leibler divergence (RKLD) appears to be the one that more closely matches the PESQ MOS-LQO scores and the Itakura divergence (ID) estimates. Ranking close behind, the Itakura-Saito divergence (ISD) comes in the CEASA assessment. As a by-product of these methods, shorter analysis windows have been found to lead to better spectral fitting even though they are not the best for synthesis and related tasks as indicated by the CEASA assessment results. Future research should focus on the conception of loss functions more suitable to the task such as perceptual losses properly adapted to the structure of the learning machine, which constrains them to be differentiable with respect to the weights. Also the measures of merit should be suitable for the specific tasks in an evolution of the CEASA assessment tool.

REFERENCES

[1] M. H. Vali and T. Bäckström, "End-to-end optimized multi-stage vector quantization of spectral envelopes for speech and audio coding," in *Proc. Interspeech*, Aug. 2021, pp. 3355–3359.

[2] M. A. Ramírez, "Hybrid autoregressive resonance estimation and density mixture formant tracking model," *IEEE Access*, vol. 6, pp. 30217–30224, 2018, doi: 10.1109/ACCESS.2018.2841802.

[3] F. Itakura and S. Saito, "Analysis-synthesis telephony based on the maximum likelihood method," in *Proc. 6th Int. Congr. Acoust.*, Y. Kohasi, Ed. Tokyo, Japan: Elsevier, Aug. 1968, pp. C-17–C-20.

[4] J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*. Berlin, Germany: Springer, 1976.

[5] K. Matsubara, T. Okamoto, R. Takashima, T. Takiguchi, T. Toda, Y. Shiga, and H. Kawai, "Full-band LPCNet: A real-time neural vocoder for 48 kHz audio with a CPU," *IEEE Access*, vol. 9, pp. 94923–94933, 2021, doi: 10.1109/ACCESS.2021.3089565.

[6] J. Lilley and H. T. Bunnell, "Unsupervised training of a DNN-based formant tracker," in *Proc. Interspeech*, Brno, Czechia, Aug. 2021, pp. 1189–1193.

[7] Y.-A. Chung and J. Glass, "Generative pre-training for speech with autoregressive predictive coding," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 3497–3501.

[8] M. A. Ramírez, "A Levinson algorithm based on an isometric transformation of Durbin's," *IEEE Signal Process. Lett.*, vol. 15, pp. 99–102, 2008, doi: 10.1109/LSP.2007.910319.

[9] Z. Cui, C. Bao, J. K. Nielsen, and M. G. Christensen, "Autoregressive parameter estimation with DNN-based pre-processing," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 6759–6763, doi: 10.1109/ICASSP40776.2020.9053755.

[10] F. Villavicencio, A. Robel, and X. Rodet, "Improving LPC spectral envelope extraction of voiced speech by true-envelope estimation," in *Proc. IEEE Int. Conf. Acoust. Speed Signal Process.*, Toulouse, France, May 2006, pp. I-869–I-873, doi: 10.1109/ICASSP.2006.1660159.

[11] Y.-J. Hu and Z.-H. Ling, "Extracting spectral features using deep autoencoders with binary distributed hidden units for statistical parametric speech synthesis," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 4, pp. 713–724, Apr. 2018.

[12] Z. H. Ling, L. Deng, and D. Yu, "Modeling spectral envelopes using restricted Boltzmann machines and deep belief networks for statistical parametric speech synthesis," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 10, pp. 2129–2139, Oct. 2013, doi: 10.1109/TASL.2013.2269291.

[13] P. Jiang, H. Fu, H. Tao, P. Lei, and L. Zhao, "Parallelized convolutional recurrent neural network with spectral features for speech emotion recognition," *IEEE Access*, vol. 7, pp. 90368–90377, 2019, doi: 10.1109/ACCESS.2019.2927384.

[14] N. Hajarolasvadi, M. A. Ramírez, W. Beccaro, and H. Demirel, "Generative adversarial networks in human emotion synthesis: A review," *IEEE Access*, vol. 8, pp. 218499–218529, 2020, doi: 10.1109/access.2020.3042328.

[15] J. Abel and T. Fingscheidt, "Artificial speech bandwidth extension using deep neural networks for wideband spectral envelope estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 1, pp. 71–83, Jan. 2018.

[16] Z.-Q. Wang, P. Wang, and D. Wang, "Complex spectral mapping for single- and multi-channel speech enhancement and robust ASR," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1778–1787, 2020.

[17] *Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs*, document ITU-T P.862, Recommendation, Feb. 2001.
[18] *Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs*, document ITU-T P.862.2, Recommendation, Nov. 2005.
[19] A. Hines, J. Skoglund, A. C. Kokaram, and N. Harte, "ViSQOL: An objective speech quality model," *EURASIP J. Audio, Speech, Music Process.*, vol. 2015, no. 1, p. 13, Dec. 2015, doi: 10.1186/s13636-015-0054-9.
[20] C. Magnant, E. Grivel, A. Giremus, L. Ratton, and B. Joseph, "Classifying autoregressive models using dissimilarity measures: A comparative study," in *Proc. 23rd Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2015, pp. 998–1002.
[21] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. Acoust. Speech Signal Process.*, vol. ASSP-23, no. 1, pp. 67–72, Feb. 1975, doi: 10.1109/TASSP.1975.1162641.
[22] H. Chen, B. Jiang, S. X. Ding, and B. Huang, "Data-driven fault diagnosis for traction systems in high-speed trains: A survey, challenges, and perspectives," *IEEE Trans. Intell. Transp. Syst.*, early access, Oct. 22, 2020, doi: 10.1109/TITS.2020.3029946.
[23] M. K. Banavar, H. Gan, B. Robistow, and A. Spanias, "Signal processing and machine learning concepts using the reflections echolocation app," in *Proc. IEEE Frontiers Educ. Conf. (FIE)*, Oct. 2017, pp. 1–5.
[24] N. S. Jayant and P. Noll, *Digital Coding of Waveforms*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1984.
[25] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall, "The DARPA speech recognition research database: Specifications and status," in *Proc. DARPA Speech Recognition Workshop*, vol. 1, Feb. 1986, pp. 93–99.
[26] E. P. Papamichalis, *Practical Approaches to Speech Coding*. Upper Saddle River, NJ, USA: Prentice-Hall, 1987.
[27] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *J. Mach. Learn. Res.*, vol. 6, no. 4, pp. 1705–1749, 2005.[Online]. Available: https://jmlr.csail.mit.edu/papers/v6/banerjee05b.html
[28] D. Ramírez, J. Vía, I. Santamaría, and P. Crespo, "Entropy and Kullback–Leibler divergence estimation based on Szegö's theorem," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, Glasgow, U.K., 2009, pp. 998–1002. [Online]. Available: https://ieeexplore.ieee.org/document/7077381
[29] A. Zhao, K. Subramani, and P. Smaragdis, "Optimizing short-time Fourier transform parameters via gradient descent," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Toronto, ON, Canada, Jun. 2021, pp. 736–740, doi: 10.1109/ICASSP39728.2021.9413704.
[30] M. C. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

MIGUEL ARJONA RAMÍREZ (Senior Member, IEEE) received the B.S. degree in electronics engineering from the Instituto Tecnológico de Aeronáutica, Brazil, in 1980, the degree in electronic design engineering from the Philips International Institute, The Netherlands, in 1981, and the M.S. and Ph.D. degrees in electrical engineering and the Habilitation degree in signal processing from the University of São Paulo, Brazil, in 1992, 1997, and 2006, respectively.

He was the Engineering Development Group Leader of Interactive Voice Response Systems (IVRs) for Itautec Informática, Brazil, where he worked, from 1982 to 1990. In 2008, he carried research in time-frequency speech analysis and coding in a research visit to the Royal Institute of Technology, Sweden. He is currently an Associate Professor with the Polytechnic School of the University of São Paulo, where he is a member of the Signal Processing Laboratory. His research interests include the application of novel signal processing and machine learning algorithms to signal compression and prediction, speech analysis, coding and recognition, speaker identification, and audio analysis and coding. He has authored or coauthored four book chapters and over 70 journals and conference papers in these areas. He is a member of the Brazilian Telecommunications Society (SBrT).

WESLEY BECCARO received the B.Sc., M.Sc., and Ph.D. degrees in electrical engineering from the University of São Paulo in 2008, 2012, and 2017, respectively. His research interests include digital signal processing, instrumentation, embedded systems, and fuel qualification.

DEMÓSTENES ZEGARRA RODRÍGUEZ (Senior Member, IEEE) received the B.S. degree in electronic engineering from the Pontifical Catholic University of Peru, and the M.S. and Ph.D. degrees from the University of São Paulo, in 2009 and 2013, respectively. He is currently an Adjunct Professor with the Department of Computer Science, Federal University of Lavras, Brazil. He has a solid knowledge in telecommunication systems and computer science based on 15 years of professional experience in major companies. In 2018 and 2019, he carried research in parametric speech quality assessment methods applied to telecommunication services in a postdoctoral position with the Technical University of Berlin, Germany. His research interests include QoS and QoE in multimedia services, architect solutions in telecommunication systems, and machine learning algorithms. He is a member of the Brazilian Telecommunications Society.

RENATA LOPES ROSA received the M.S. degree from the University of São Paulo, in 2009, and the Ph.D. degree from the Polytechnic School of the University of São Paulo (EPUSP), in 2015. She is currently an Adjunct Professor with the Department of Computer Science, Federal University of Lavras, Brazil. She has a solid knowledge in computer science based on more than ten years of professional experience. Her current research interests include computer networks, telecommunication systems, machine learning, quality of experience of multimedia service, social networks, and recommendation systems.