Speech enhancement by low-rank and convolutive dictionary spectrogram decomposition

Zhuo Chen[1,2], Brian McFee[1], Daniel P. W. Ellis[1,2]

[1]LabROSA, Columbia University, New York, NY, USA
[2]International Computer Science Institute, Berkeley, CA, USA

zc2204@columbia.edu

Abstract

A successful speech enhancement system requires strong models for both speech and noise to decompose a mixture into the most likely combination. However, if the noise encountered differs significantly from the system's assumptions, performance will suffer. In previous work, we proposed a speech enhancement framework based on decomposing the noisy spectrogram into low rank background noise and a sparse activation of pre-learned templates, which requires few assumptions about the noise and showed promising results. However, when the noise is highly non-stationary or has large amplitude, the local SNR of the noisy speech can change drastically, resulting in less accurate decompositions between foreground speech and background noise. In this work, we extend the previous model by changing the modeling of the speech part to be the convolution of a sparse activation and pre-learned template patches, which enforces continuous structure within the speech and leads to better results in highly corrupted noisy mixtures.

Index Terms: speech enhancement, convolutive factorization, patch dictionary

1. Introduction

Automatically enhancing degraded and noisy speech is one of the key problems in speech processing. Speech enhancement serves both to pre-process audio for automatic speech recognition, as well as improve quality for human listeners. In general speech enhancement, the signal is typically modeled as a combination of a clean speech with a noisy background, and the goal is to recover the speech component by detecting and suppressing noise. Consequently, to ensure high-quality speech enhancement, both the estimation of speech and the noise need to be accurate. However, since "noise" can vary widely from one context to the next, it can be difficult to formulate a single noise model which performs well in all situations.

Existing speech enhancement systems typically make several assumptions about the noise distribution. In the traditional speech enhancement framework, noise is assumed to be stationary and to follow a Gaussian distribution across each frequency bin of the spectrogram, with parameters that can be estimated from detected gaps in the speech [1]. To deal with non-stationary noise, several fixed-rank noise models have been proposed, where the noise is assumed to lie in the span of a low-rank (non-negative) subspace [2, 3]. But when noise fails to match the assumptions of the model, enhancement quality rapidly declines.

The recently-developed technique of Robust Principal Component Analysis (RPCA) [4] provides a new approach to

---

distinguishing background noise. RPCA decomposes a matrix into the sum of a sparse component and a low-rank component. Some speech enhancement scenarios include background noise that can be non-stationary but still exhibit low-rank structure. Noise is often less spectrally diverse than foreground speech, in which case it will be captured by the low-rank component when RPCA is applied to the spectrogram.

In recent work, we proposed a model that further decomposes the sparse component of RPCA into the product of a pre-learned dictionary of spectra and a sparse activation matrix, and where the background noise is modeled as the sum of a low-rank matrix and a Gaussian residual [5]. The key feature of the RPCA formulation is that the rank of the noise basis is inferred from the input signal, rather than fixed in advance.

However, as described above, the effectiveness of speech enhancement depends on both the noise and speech models. One drawback of the previous model [5] was that the speech spectrogram was represented as frame-wise sparse activations of pre-learnt spectra which did not explicitly account for temporal continuity of speech. Consequently, the performance of the model degrades in low SNR situations or when the noise is highly transient.

In this work, we incorporate the temporal continuity of speech by extending the previous model [5] to decompose the noisy spectrogram into the sum of two components: a low rank background noise matrix, and the convolution of pre-learned time-varying templates with a sparse activations. Sparse convolutional modeling of speech allows us to discover a compact representation of the audio which exploits continuous structure in speech. This results in improved robustness, and better performance in moderate-to-low SNRs.

2. Proposed Model

2.1. Background: Low-rank noise

Candès *et al.* showed that under broad conditions, a data matrix $Y \in \mathbb{R}^{F \times T}$ can be uniquely decomposed as the sum of sparse matrix S and low-rank matrix L by solving the following convex optimization problem [4]:

$$\min_{S,L} \|S\|_1 + \|L\|_* \qquad \text{s.t.} \qquad Y = S + L. \qquad (1)$$

In (1), $\|S\|_1$ denotes the element-wise ℓ_1-norm, which provides a convex surrogate for sparsity of S. $\|L\|_*$ denotes the nuclear norm — or sum of singular values — which provides a convex surrogate for the rank of L. Because of its ability to reveal the intrinsic structure of a matrix, RPCA has been shown to have good performance in the problem of audio source separation,

where S and L are taken as the foreground and the background components in the mixture [6].

For speech enhancement, where the target is to extract clean speech, the objective is more specific than in general source separation. To better constrain the speech model, we previously adapted RPCA by replacing the sparse component of the RPCA with a sparse activation $H \in \mathbb{R}_+^{K \times T}$ of a pre-learned speech codebook $W \in \mathbb{R}_+^{F \times K}$ [5]:

$$\min_{H,L,E} \|E\|_{\mathrm{F}}^2 + \lambda_H \|H\|_1 + \lambda_L \|L\|_* + \mathcal{I}_+(H) \tag{2}$$
$$\text{s.t.} \qquad Y = WH + L + E.$$

Here, $\|E\|_{\mathrm{F}}^2$ denotes the Frobenius norm of the residual E. The indicator function $\mathcal{I}_+(H)$ constrains the activation H to be non-negative, which prevents the energy cancellation with the negative weight on the codewords.

2.2. Convolutive sparse low-rank decomposition

One problem of the column-wise model described above, as well as other column-based factorization systems [7, 8, 2], is that they assume that frames are invariant to temporal permutations. However, speech clearly exhibits strong temporal dependencies which can be exploited to improve enhancement [1].

To better capture temporal dynamics of speech, we propose a convolutive extension of (2), which takes the form:

$$\min_{H,L,E} \|E\|_{\mathrm{F}}^2 + \lambda_H \|H\|_1 + \lambda_L \|L\|_* + \mathcal{I}_+(H) \tag{3}$$
$$\text{s.t.} \qquad Y = \sum_{\tau=0}^{P-1} W(\tau) \overset{\tau \rightarrow}{H} + L + E.$$

Here, $\{W(\tau)\} \subset \mathbb{R}_+^{F \times K}, \tau = 1, ..., P$ is a set of time-varying basis elements, where each $W(\tau)$ encodes the spectra pattern of each patch at its τth frame. $H \in \mathbb{R}_+^{K \times T}$ is the corresponding set of non-negative convolutive activations, and $\overset{\tau \rightarrow}{H}$ refers the "shift" operation, which pads τ zero-columns to the left of H and truncates its rightmost $P - \tau$ columns to maintain shape, with $\overset{\leftarrow \tau}{H}$ defined analogously for left-shift.

Compared with (2), the proposed model decomposes speech as the sum of convolutions between the dictionary elements and their corresponding activations. Rather than individual speech spectra, the dictionary now consists of two-dimensional "patches" of speech, which capture the energy distribution in each frequency bin over subsequent points in time. Modeling temporal dependencies in this way prevents the speech model from erroneously capturing transient noise bursts. The proposed system thus provides robust estimation of both speech and noise components.

The parameter P controls the length of convolution window. By controlling the convolution window length, parameter P affects the model's emphasis on continuity. Note that when $P = 1$, the proposed model reduces to the column based model. Larger values of P correspond to longer basis patches, and enforce longer dynamic structure within the estimated speech. P therefore provides a parameter to trade-off between flexibility and reconstruction accuracy. In low SNR situations, longer convolution windows prevent the system from modeling noise with the speech components. Conversely, when SNR is relatively high, enforcing long dynamic dependencies in the speech would limit the model in representing the detail of the speech, and lead to inaccurate reconstruction.

Algorithm 1 Convolutive sparse low-rank decomposition

Input: noise+speech spectrogram Y, convolutive basis W,
Output: activations H, noise spectrogram L
Initialization: $H \leftarrow$ random positive values; $L \leftarrow \mathbf{0}$
for $t = 1, 2, \ldots$ until convergence **do**
    update H:
      $R \leftarrow \left(Y - L^t\right)_+$
      $Z \leftarrow \sum_\tau W(\tau) \overset{\tau \rightarrow}{H^t}$
      $H_\tau \leftarrow H^t \circ \dfrac{\left(W(\tau)^{\mathsf{T}} \overset{\leftarrow \tau}{R} - \lambda_H \mathbf{1}^{K \times T}\right)_+}{W(\tau)^{\mathsf{T}} \overset{\leftarrow \tau}{Z}}$
      $H^{t+1} \leftarrow \frac{1}{T} \sum_\tau H_\tau$
    update L:
      $U, \Sigma, V^{\mathsf{T}} \leftarrow \text{svd}\left(Y - \sum_\tau W(\tau) \overset{\tau \rightarrow}{H^{t+1}}\right)$
      $L^{t+1} \leftarrow U \mathcal{S}_{\lambda_L}(\Sigma) V^{\mathsf{T}}$
end for

The proposed model can also be viewed as a natural extension of convolutive non-negative matrix factorization (CNMF). In the traditional CNMF/NMF model for speech enhancement, the noise spectrogram Y_{n} was modeled as a fixed rank matrix $Y_{\mathrm{n}} = W_{\mathrm{n}} H_{\mathrm{n}}$, where the size of the noise dictionary must be set beforehand [9, 3]. However, the appropriate rank for the noise varies with SNR and noise types, and a single fixed-rank model may not work in all situations. Moreover, fixed-rank noise models need to estimate both the noise dictionary and the corresponding activation, which results in a difficult, non-convex optimization problem. By replacing the fixed-rank noise model with an adaptive low-rank penalty, the model can adjust its complexity in modeling noise for each input mixture. Since the nuclear norm optimization is convex, a global optimum can be guaranteed.

2.3. Algorithm

Given the dictionary W, the problem given in (3) is jointly convex in L and H, and a global optimum can be obtained by a variety of methods, *e.g.*, projected sub-gradient descent. Here, we opt for alternating block optimization, primarily because it leads to simple and efficient multiplicative updates, and the resulting algorithm is qualitatively similar to those for (convolutive) NMF [10, 9]. Note that the variable E may be eliminated by substituting the constraint into the objective.

When updating H, the current L is held constant, which makes the model equivalent to CNMF applied to $Y - L$. We can then compute the element-wise gradient of (3) with respect to H to obtain multiplicative update rules for H. By initializing H with positive values, the non-negativity of H can be ensured.

When updating L, H is held constant, and the resulting update is the standard nuclear norm proximal problem [11]. The optimum can be computed through soft thresholding the singular values of L.

The full algorithm is detailed in Algorithm 2.3, where $\circ$ denotes Hadamard multiplication, $\mathbf{1}^{K \times T}$ denotes the $K \times T$ all-ones matrix, $(\cdot)_+$ denotes projection onto the non-negative orthant, and $\mathcal{S}_\lambda$ denotes the soft-threshold operator:

$$\mathcal{S}_\lambda(x) := \begin{cases} \left(1 - \frac{\lambda}{|x|}\right) x & \text{if } |x| > \lambda \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Note that by fixing $L = 0$, we recover the original sparse

CNMF model [9]. For a fixed H, the resulting multiplicative update rule for W is

$$W(\tau) \leftarrow W(\tau) \circ \frac{\overset{\tau \rightarrow}{Y H^{\mathsf{T}}}}{\overset{\tau \rightarrow}{Z H^{\mathsf{T}}}}. \qquad (5)$$

Combining these facts allows us to learn a codebook W from examples of clean speech by performing alternating minimization of W and H.

3. Experiments

The proposed system was evaluated with 2500 noisy speech examples, totaling 2.5 hours. The noisy signals were synthesized by adding clean speech to a variety of noise signals at different SNRs. Clean speech was randomly sampled from the TIMIT dataset [12]. Noise data was drawn from AURORA dataset [13] and the collection used in by Duan *et al.* [7]. We include 8 stationary noise types: *car*, *exhibition*, *restaurant*, *babble*, *train*, *subway*, *train*, *airport*; and two transient noises: *keyboard* and *birds*. Test samples were mixed with noise at four SNRs, ranging from -10 to 5dB. All signals were resampled to 8kHz, and spectrograms were calculated using a window of 32ms and a hop of 10ms.

The speech dictionary was learned from 200 utterances of 20 speakers, which were disjoint from the speakers used to make the test samples. For each speaker, a dictionary of 50 elements was constructed by alternating minimization of (3) and (5). The final dictionary was formed by concatenating each speaker-dependent dictionary, resulting in $K = 1000$ basis elements. To measure the influence of window length, we evaluated the model using two convolution window lengths: $P = 3$ and $P = 5$. The weights λ_L and λ_H were tuned on a small held-out set to values 0.08 and 5.5.

Five models were selected as baselines for comparison. The first three were sparse convolutive NMF with rank 1, 4 and 8 (FR1, FR4, and FR8, respectively). We also include the previous column-based low-rank noise model (2) as a baseline (LR). To ensure fair comparison, dictionaries for the column-based systems were learned from the same training set with the same size as the total number of frames in the convolutive dictionary. We also include the classic LogMMSE estimation [1], using the implementation provided in [14].

We evaluated speech enhancement using two metrics. We used the popular BSS_EVAL package [15] to calculate the Signal-to-Distortion Ratio (SDR). And The second criteria was the PESQ estimate of subjective speech quality[16]. For both metrics, a larger score indicates better performance. The results are shown in Table 1–3 and Figures 1&2.

4. Results and Discussion

As illustrated in Table 1 & 2, the proposed CLR algorithm achieves the highest performance in all SNR conditions, except against FR1 at the 0dB SNR. And for PESQ score, the CLR outperforms all the base-line except the 5 dB experiment, in which logMMSE achieves the highest score.

In low-SNR conditions, the noise has large amplitude, and correspondingly large variance. This effectively forces the speech model to represent the large noise energy in spectrogram in order to reduce the reconstruction error. Increasing the rank of the noise model can help to mitigate this effect, as it provides greater model flexibility to separate noise from speech. This can be observed in the -10dB experiment, where the rank-4

SNR	FR1	FR4	FR8	LR	LogMMSE	CLR
-10dB	-2.98	-2.26	-2.07	-4.13	-5.62	**-0.86**
-5dB	3.03	2.15	1.02	1.47	0.31	**3.63**
0dB	**7.39**	4.32	1.55	5.79	5.18	7.22
$+5$dB	**9.39**	4.56	1.11	8.44	9.28	**9.33**

Table 1: SDR values (in dB) for all systems at various SNRs, averaged across all noise types. FR1, FR4 and FR8 are sparse convolutive NMF models with rank 1, 4 and 8. LR corresponds to (2) [5]. CLR is the proposed convolutive low-rank model($P = 5$). For each SNR, bold indicates statistical equivalence to the best result under a Bonferroni-corrected Wilcoxon sign-rank test against CLR at sensitivity $\alpha = 0.05$.

SNR	FR1	FR4	FR8	LR	LogMMSE	CLR
-10dB	1.35	1.35	1.32	0.93	1.16	**1.43**
-5dB	1.67	1.58	1.47	1.17	1.54	**1.71**
0dB	1.97	1.73	1.47	1.42	**1.97**	**1.98**
$+5$dB	2.16	1.76	1.38	1.62	**2.36**	2.22

Table 2: PESQ scores for all systems at various SNRs, averaged across all noise types.

and rank-8 models outperform the rank-1 model. Conversely, in higher SNR conditions, the noise is relatively small, and a rank-1 noise model suffice. In the experiments at -5dB and above, the trend is reversed, and FR1 outperforms FR4 and FR8. However, if the fixed rank is too high, as in FR4 and FR8, the extra representation power would erroneously capture the speech energy, and impair separation performance.

In high-SNR conditions (5dB and above), the speech model itself can generate the most robust enhancement result, as even rank-1 approximation would suffer the energy loss problem. In general, the complexity of the noise model should be adaptive: it must be rich enough to capture background noise, but not so rich as to capture speech.

A similar analysis can also be performed for each particular type of noise. Figure 1 depicts the per-noise enhancement results for four typical noises. We observe that under a fixed SNR, the performance of different systems varies significantly according to the specific type of noise in the mixture. For transient noises, such as *bird* and *keyboard*, a high-rank noise model is necessary to prevent the speech model from fitting noise. For relatively stationary noises, such as *train* and *car*, energy loss could be the main reason for the unsuccessful separation, and low-rank noise models suffice.

As discussed in section 2, the proposed model can adaptively determine the noise model complexity according to the balance between the speech and noise energy. Therefore, in low-SNR and transient noise experiments, the proposed model should adapt to a high rank noise mode (sometimes higher than 8), and thus generate better or equivalent results as FR4 and FR8. On the other hand, for stationary noise or high-SNR settings, where FR4 and FR8 suffered from the energy loss problem, the proposed model should automatically decrease its noise rank (zero for many samples in 5dB SNR experiments). Note that in some experiment settings, the proposed model also slightly suffers from the incorrect noise rank problem, resulting in lower SDR than FR1 in 0dB experiment, where the optimum rank for most noises was one. However, for a large variety of SNR conditions, the proposed model provides significantly more robust separation, which makes it especially suitable for real-world applications when the SNR and noise types are unknown.
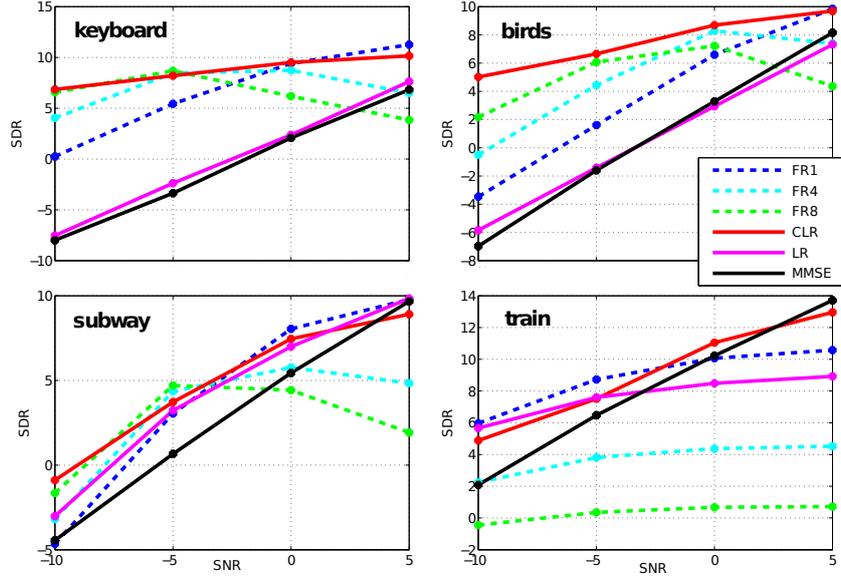
Figure 1: Enhancement results for various types of noise.

	SDR			PESQ		
SNR	LR	CLR3	CLR5	LR	CLR3	CLR5
−10dB	-3.65	-2.06	**-0.64**	0.93	1.30	**1.43**
−5dB	1.80	2.72	**3.93**	1.17	1.59	**1.71**
0dB	5.94	6.89	**7.32**	1.43	1.90	**1.98**
+5dB	9.35	**10.32**	9.25	1.63	**2.24**	2.22

Table 3: SDR values and PESQ scores for each convolutive window size.

Table 3 shows the performance of the proposed model with $P = 1, 3$ and 5, where $P = 1$ corresponds to the column-based system. As discussed in section 2, the longer window enforces stronger continuity constraints, leading to better performance of CLR in low-SNR and transient noise experiments. As the SNR increases, noisy speech was easier to separate, which reduces the necessity for continuity, but calls for stronger modeling power to accurately reconstruct the speech. In those conditions, the short-window model (CLR3) tends to outperform the long window model (CLR5), as observed in the +5dB experiment in table 3. And for the same reason, the PESQ score of the CLR5 , which emphasizes more in the speech quality than the noise reduction, was lower than LogMMSE and CLR3 when the SNR is at +5dB.

5. Conclusion

In this work, we proposed an algorithm that decomposes noisy speech into the sum of a low rank matrix (capturing background noise with limited spectral variation) and the convolution between a learned basis of speech patches and corresponding sparse activations. When employed for speech enhancement, the proposed model achieves significantly better performance in transient noise and moderate-to-low SNR conditions than fixed-rank factorizations and non-convolutive models, .
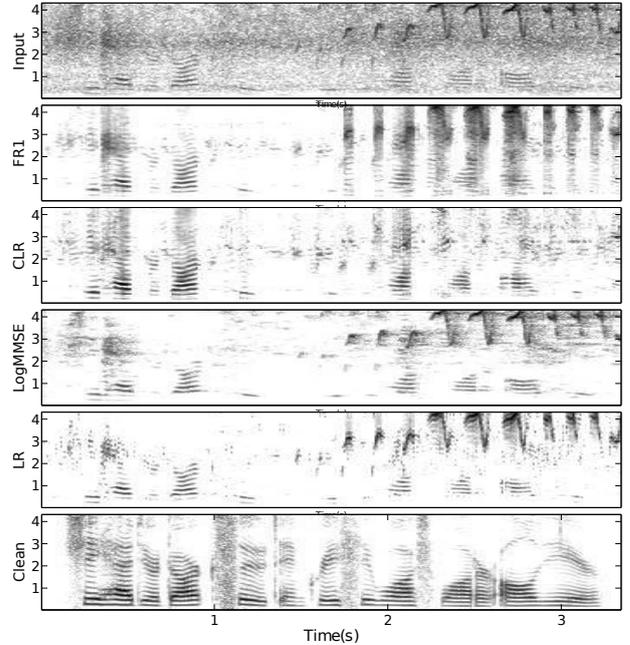

Figure 2: Example decomposition. The top pane shows the spectrogram of speech mixed with bird chirping at −5dB. The output of FR1 is shown in the second pane, and the third pane is the result for CLR3. The fourth and fifth pane show log-MMSE and LR enhancement. The clean speech appears in the bottom pane.

6. References

[1] E. Yariv and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust. Speech Signal Process*, vol. 33, pp. 443–445, 1985.

[2] D. L. Sun and G. J. Mysore, "Universal speech models for speaker independent single channel source separation," *Proc. IEEE ICASSP*, 2013.

[3] M. Carlin, N. Malyska, and T. F. Quatieri, "Speech enhancement using sparse convolutive non-negative matrix factorization with basis adaptation," *Proceedings of Interspeech*, 2012.

[4] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *arXiv preprint arXiv:0912.3599*, 2009.

[5] Z. Chen and D. P. Ellis, "Speech enhancement by sparse, low-rank, and dictionary spectrogram decomposition," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 1–4, 2013.

[6] P.-S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *Proc. IEEE ICASSP*, 2012, pp. 57–60.

[7] Z. Duan, G. J. Mysore, and P. Smaragdis, "Speech enhancement by online non-negative spectrogram decomposition in non-stationary noise environments," *Proceedings of Interspeech*, 2012.

[8] M. Schmidt and O. Rasmus, "Single-channel speech separation using sparse non-negative matrix factorization," *Proceedings of Interspeech*, p. 26142617, 2006.

[9] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 1–12, 2007.

[10] A. Cichocki, R. Zdunek, and S.-i. Amari, "New algorithms for non-negative matrix factorization in applications to blind source separation," *Proc. IEEE ICASSP*, vol. 5, 2006.

[11] N. Parikh and S. Boyd, "Proximal algorithms," *Foundations and Trends in optimization*, vol. 1, pp. 123–231, 2013.

[12] J. S. Garofolo, L. D. Consortium *et al.*, *TIMIT: acoustic-phonetic continuous speech corpus*. Linguistic Data Consortium, 1993.

[13] H.-G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*, 2000.

[14] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Taylor and Francis, 2007.

[15] E. Vincent, R. Gribonval, and C. Fvotte, "Performance measurement in blind audio source separation," *IEEE Trans. on Audio Speech Lang. Process*, vol. 14, pp. 1462–1469, 2006.

[16] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality - a new method for speech quality assessment of telephone networks and codes," in *Proc. IEEE ICASSP*, 2001, pp. 749–752.