Enhancing Sparsity in Linear Prediction of Speech by Iteratively Reweighted 1-norm Minimization

Giacobello, Daniele; Christensen, Mads Græsbøll; Murthi, Manohar N.; Jensen, Søren Holdt; Moonen, Marc

Published in:
I E E E International Conference on Acoustics, Speech and Signal Processing. Proceedings

DOI (link to publication from Publisher):
10.1109/ICASSP.2010.5495198

Publication date:
2010

Document Version
Accepted author manuscript, peer reviewed version

Citation for published version (APA):
Giacobello, D., Christensen, M. G., Murthi, M. N., Jensen, S. H., & Moonen, M. (2010). Enhancing Sparsity in Linear Prediction of Speech by Iteratively Reweighted 1-norm Minimization. *I E E E International Conference on Acoustics, Speech and Signal Processing. Proceedings*, *2010*, 4650-4653 . https://doi.org/10.1109/ICASSP.2010.5495198

ENHANCING SPARSITY IN LINEAR PREDICTION OF SPEECH BY ITERATIVELY REWEIGHTED 1-NORM MINIMIZATION

Daniele Giacobello[1], *Mads Græsbøll Christensen*[1], *Manohar N. Murthi*[2],
Søren Holdt Jensen[1], *Marc Moonen*[3]

[1]Dept. of Electronic Systems, Aalborg Universitet, Denmark
[2]Dept. of Electrical and Computer Engineering, University of Miami, USA
[3]Dept. of Electrical Engineering (ESAT-SCD), Katholieke Universiteit Leuven, Belgium
{dg,mgc,shj}@es.aau.dk, mmurthi@miami.edu, marc.moonen@esat.kuleuven.be

ABSTRACT

Linear prediction of speech based on 1-norm minimization has already proved to be an interesting alternative to 2-norm minimization. In particular, choosing the 1-norm as a convex relaxation of the 0-norm, the corresponding linear prediction model offers a sparser residual better suited for coding applications. In this paper, we propose a new speech modeling technique based on reweighted 1-norm minimization. The purpose of the reweighted scheme is to overcome the mismatch between 0-norm minimization and 1-norm minimization while keeping the problem solvable with convex estimation tools. Experimental results prove the effectiveness of the reweighted 1-norm minimization, offering better coding properties compared to 1-norm minimization.

Index Terms— Linear prediction, 1-norm minimization, speech analysis, speech coding.

1. INTRODUCTION

In Linear Predictive Coding of speech signals (LPC), the prediction coefficients are typically obtained by minimizing the 2-norm of the residual (the difference between the observed signal and the predicted signal) [1]. The 2-norm minimization shapes the residual into variables that exhibit Gaussian-like characteristics. However, in order to reduce the information content of the residual and to allow for a low bit rate encoding, a sparse approximation of the residual is often used. This conceptual difference between a quasi-white minimum variance residual and its approximated version creates a mismatch that can raise the distortion significantly. In our recent work, we have defined a new predictive framework that provides a tighter coupling between the linear predictive analysis and the residual encoding by looking for a sparse residual rather than a minimum variance one [2, 3]. Early encoding techniques such as Multi-Pulse Excitation (MPE) [4] or Regular-Pulse Excitation (RPE) [5], have shown to be more consistent with this kind of predictive framework unlike, e.g., Code Excited LP (CELP) [6] that uses pseudo-random sequences to encode the residual.

In our previous work we have used the 1-norm as a convex relaxation of the so-called 0-norm, the cardinality of a vector. The 0-norm, and more generally the p-norm with $0 \leq p < 1$, is not a proper norm and its minimization yields a combinatorial problem (NP-hard). We therefore aim to "adjust" the error weighting difference between the 1-norm and the 0-norm keeping the feasibility of the problem in polynomial time. To do so, in this paper we propose a new method for the estimation of the prediction filter based on iteratively reweighted 1-norm minimization [7]. We will see how this method, by enhancing the sparsity of the residual, yields a better and simpler formulation of the coding problem, hence allowing for a general improvement in performance.

The paper is organized as follows. In Section 2 we give the general problem formulation of sparse linear prediction. In Section 3 we introduce the algorithms used to enhance sparsity in linear predictive coding and in Section 4 we provide a statistical interpretation. In Section 5 and Section 6 we illustrate the effects of the algorithm for analysis and coding of speech. Section 7 concludes the paper.

2. SPARSE LINEAR PREDICTION

The problem considered in this paper is based on the following Auto-Regressive (AR) speech production model, where a sample of speech $x(n)$ is written as a linear combination of K past samples:

$$x(n) = \sum_{k=1}^{K} a_k x(n-k) + r(n), \quad 0 < n \leq N, \tag{1}$$

where $\{a_k\}$ are the prediction coefficients and $r(n)$ is the driving noise process (commonly referred to as the prediction residual). The speech production model (1) in matrix form becomes:

$$\mathbf{x} = \mathbf{X}\mathbf{a} + \mathbf{r} \tag{2}$$

where:

$$\mathbf{x} = \begin{bmatrix} x(N_1) \\ \vdots \\ x(N_2) \end{bmatrix}, \mathbf{X} = \begin{bmatrix} x(N_1-1) & \cdots & x(N_1-K) \\ \vdots & & \vdots \\ x(N_2-1) & \cdots & x(N_2-K) \end{bmatrix}.$$

The prediction coefficient vector $\mathbf{a} \in \mathbb{R}^K$ is found by minimizing the $p-$norm of the residual $\mathbf{r}$ [8]:

$$\hat{\mathbf{a}}, \hat{\mathbf{r}} = \arg\min_{\mathbf{a}} \|\mathbf{r}\|_p^p, \quad \text{s.t.} \quad \mathbf{r} = \mathbf{x} - \mathbf{X}\mathbf{a}; \tag{3}$$

where $\|\cdot\|_p$ is the p-norm. The starting and ending points $N_1 = 1$ and $N_2 = N + K$ are chosen assuming that $x(n) = 0$ for $n < 1$ and $n > N$ [9]. Sparsity is often measured as the cardinality, i.e.,

The work of Daniele Giacobello is supported by the Marie Curie EST-SIGNAL Fellowship (http://est-signal.i3s.unice.fr), contract no. MEST-CT-2005-021175.

The work of Manohar N. Murthi is supported by the National Science Foundation via awards CCF-0347229 and CNS-0519933.

Algorithm 1 Iteratively Reweighted 1-norm Minimization of the Residual

```
Inputs: speech segment $\mathbf{x}$
Outputs: predictor $\hat{\mathbf{a}}^i$, residual $\hat{\mathbf{r}}^i$
$i = 0$, initial weights $\mathbf{W}^{i=0} = \mathbf{I}$
while halting criterion false do
    1. $\hat{\mathbf{a}}^i, \hat{\mathbf{r}}^i \leftarrow \arg\min_{\mathbf{a}} \|\mathbf{W}^i\mathbf{r}\|_1 \quad \text{s.t.} \quad \mathbf{r} = \mathbf{x} - \mathbf{X}\mathbf{a}$
    2. $\mathbf{W}^{i+1} \leftarrow \text{diag}\left(|\hat{\mathbf{r}}^i| + \epsilon\right)^{-1}$
    3. $i \leftarrow i + 1$
end while
```

the so-called 0-norm. Therefore, setting $p = 0$ in (3) means that we aim to minimize the number of non-zero samples in the error signal. Unfortunately this corresponds to a combinatorial problem which generally cannot be solved in polynomial time. Instead of the 0-norm, we then use the more tractable 1-norm [2]:

$$\hat{\mathbf{a}}, \hat{\mathbf{r}} = \arg\min_{\mathbf{a}} \|\mathbf{r}\|_1 \quad \text{s.t.} \quad \mathbf{r} = \mathbf{x} - \mathbf{X}\mathbf{a}; \tag{4}$$

An interesting alternative problem formulation is obtained when sparsity is also imposed on the predictor:

$$\hat{\mathbf{a}}, \hat{\mathbf{r}} = \arg\min_{\mathbf{a}} \|\mathbf{r}\|_1 + \gamma\|\mathbf{a}\|_1, \quad \text{s.t.} \quad \mathbf{r} = \mathbf{x} - \mathbf{X}\mathbf{a}; \tag{5}$$

in this case the sparse structure of the predictor (in this case high order) allows a joint estimation of a short-term and a long-term predictor [3, 10]. This optimization problem can be posed as a linear programming problem and can be solved using an interior-point algorithm [8].

3. ITERATIVELY REWEIGHTED 1-NORM MINIMIZATION

Our general goal is to determine a linear predictor that yields a sparse residual. As mentioned before, for $0 \leq p < 1$, the problem cannot be solved using convex optimization. To overcome this problem, an iteratively reweighted 1-norm minimization may be used for estimating $\mathbf{a}$ and enhancing the sparsity on $\mathbf{r}$, while keeping the problem solvable with convex tools [7]. The algorithm is shown in Algorithm 1. The parameter $\epsilon > 0$ is used to provide stability when a component of $\hat{\mathbf{r}}$ goes to zero. ϵ does not need to be too small; as empirically demonstrated in [7], it should be in the order of the expected nonzero magnitude of $\mathbf{r}$. It can be shown that $\|\hat{\mathbf{r}}^{i+1}\|_1 \leq \|\hat{\mathbf{r}}^i\|_1$, meaning that this is a descent algorithm [7]. The halting criterion can therefore be chosen as either a maximum number of iterations or as a convergence criterion.

When we impose sparsity both on the residual and on the predictor, as in (5), the algorithm is modified as shown in Algorithm 2. As mentioned before, the high order sparse predictor estimated in (5) is found to show a structure similar to the convolution between a short-term and a long-term predictor, usually estimated in two different stages. In previous approaches [3, 10], the predictor shows a clear sparse structure but also some spurious components, i.e., small components in the predictor that are irrelevant to our analysis. In [3], we have used a model order selection criterion to locate the spurious quasi-zero components in the predictor which are then put to zero. The reweighted 1-norm minimization seems to be more effective in removing these spurious components, as the new predictor is iteratively re-estimated, rather than just "cleaned up".

Algorithm 2 Iteratively Reweighted 1-norm Minimization of Residual and Predictor

```
Inputs: speech segment $\mathbf{x}$
Outputs: predictor $\hat{\mathbf{a}}^i$, residual $\hat{\mathbf{r}}^i$
$i = 0$, initial weights $\mathbf{W}^{i=0} = \mathbf{I}$ and $\mathbf{D}^{i=0} = \mathbf{I}$
while halting criterion false do
    1. $\hat{\mathbf{a}}^i, \hat{\mathbf{r}}^i \leftarrow \arg\min_{\mathbf{a}} \|\mathbf{W}^i\mathbf{r}\|_1 + \gamma\|\mathbf{D}^i\mathbf{a}\|_1$
                    $\text{s.t.} \quad \mathbf{r} = \mathbf{x} - \mathbf{X}\mathbf{a}$
    2. $\mathbf{W}^{i+1} \leftarrow \text{diag}\left(|\hat{\mathbf{r}}^i| + \epsilon\right)^{-1}$
    3. $\mathbf{D}^{i+1} \leftarrow \text{diag}\left(|\hat{\mathbf{a}}^i| + \epsilon\right)^{-1}$
    4. $i \leftarrow i + 1$
end while
```

4. STATISTICAL INTERPRETATION

The linear prediction solution defined in (4) and (5) can be seen respectively as the *Maximum Likelihood* (ML) and *Maximum A Priori* (MAP) estimate of an AR process driven by a Laplacian noise sequence $\mathbf{r}$. In the MAP approach, a prior on $\mathbf{a}$ as a Laplacian variable is also imposed. The Laplacian distribution has already been considered to provide a more appropriate fitting for speech [12] than the Gaussian distribution, due to the heavier tails that admit larger errors in the residual. For the case $p \leq 1$, the density functions will have even heavier tails and a sharper slope near zero. In particular, this means that the maximization will encourage small values to become smaller while leaving unchange the larger values. The limit case for $p = 0$ will have an infinitely sharp slope in zero and equally weighted larger slopes. This will force the maximization to include as many zeros as possible as they are infinitely weighted.

The mismatch between the 0-norm and the 1-norm minimization that we are trying to compensate for, can be seen more clearly in Figure 1, where larger coefficients are penalized more heavily by the 1-norm than small ones. In this sense, the 0-norm can be seen as more "impartial" by penalizing every nonzero coefficient equally. It is clear that if a very small value would be weighted as much as a large value, the minimization process will try to eliminate the smaller ones and enhance the larger ones.

This explains the choice of the weights as the inverse of the magnitude of the residual. In fact, this weighting will balance the dependence on the magnitude of the 1-norm, changing the cost function and moving the problem towards the 0-norm minimization.

5. EXPERIMENTAL ANALYSIS

To illustrate the effects of the algorithm, we first analyze a segment of stationary voiced speech. The reweighted 1-norm minimization helps to reduce the emphasis on the outliers due to the pitch excitation, as we can see clearly in Figure 2. The ability of easily spotting the main components in the residual, as we shall see in the next section, have a great impact on coding applications.

An even more interesting case, is the reweighted 1-norm minimization of both residual and predictor. In this case, the use of the high order predictor removes also the long-term redundancies, what is left is almost just an impulse as shown in Figure 3. This basically means that all the information of the signal is transferred to the predictor which also show a very clear sparse structure, similar to the convolution between the coefficients of short-term and long-term predictors. The examples were obtained analyzing the vowel /a/ uttered by a female speaker using $N = 160$, $f_s = 8$ kHz and order $K = 10$ for Algorithm 1 and $K = 110$ for Algorithm 2. In

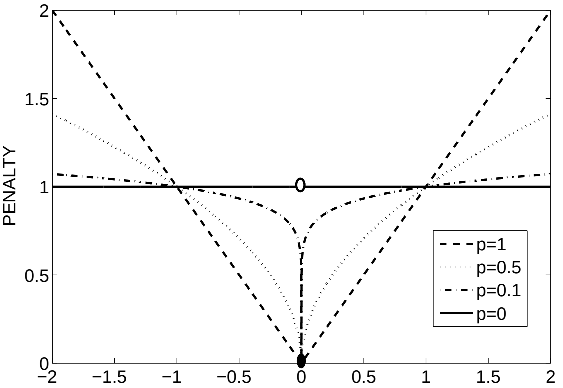

Fig. 1. *Comparison between cost functions for $p \leq 1$. The 0-norm can be seen as more "democratic" than any other norm by weighting all the nonzero coefficients equally.*

both cases $\epsilon = 0.01$. The choice of the regularization term γ is given by the L-curve where a trade-off between the sparsity of the residual and the sparsity of the predictor is found [11, 3]. Both algorithms converge rapidly, three to five iteration are sufficient to reach a point where $\|\hat{\mathbf{r}}^{i+1}\|_1 \approx \|\hat{\mathbf{r}}^{i}\|_1$ and, in the joint case, $\|\hat{\mathbf{a}}^{i+1}\|_1 \approx \|\hat{\mathbf{a}}^{i}\|_1$.

6. VALIDATION

To validate our method, we have analyzed about one hour of clean speech coming from several different speakers with different characteristics (gender, age, pitch, regional accent) taken from the TIMIT database, re-sampled at 8 kHz. The frame length is $N = 160$ (20 ms). We will consider now the two cases, with the reweighted minimization of the residual and with the reweighted minimization of both residual and predictor. The parameter ϵ, used to avoid division by zero, is chosen to be $\epsilon = 0.01$.

6.1. Reweighted Residual

In order to code the residual sequence when Algorithm 1 is used, after the reweighted scheme we use an Analysis-by-Synthesis to optimize the amplitudes of the $M = 20$ largest pulses (therefore constraining the positions). The order of the predictor is $K = 10$, a long-term predictor is not used for immediacy of the results. Our method (**MPE1r**) is compared with the classic MPE scheme where the linear predictor is found with a 1-norm minimization (**MPE1**), with a 2-norm minimization (**MPE2r**) [4] and using a 2-norm reweighted minimization (**MPE1r**) [13]. In the reweighted cases, five iterations are done (enough to reach reasonable convergence). The quantization process uses 20 bits to encode the predictor using 10 Line Spectral Frequencies using the procedure in [14], in the case the filter is unstable the poles outside the unit circle are reflected inside of it. A 3 bits uniform quantizer that goes from the lowest to the highest magnitude of the residual pulses is used to code the residual, 5 bits are used to code the lowest magnitude and 2 bits are used to code the difference between lowest and highest magnitude. The signs are coded with 1 bit per each pulse. We postpone the efficient encoding of the positions to further investigation, for now we just use the information content of the pulse location which is $\log_2 \binom{160}{10}$ bits. This produces a bit rate of 9500 bits/s. The results are shown in Table 1. We would like to highlight that in **MPE1r** it is not necessary to calculate the positions of the nonzero pulses are located (as it is usually done in MPE coding), we simply exploit the information coming out of the predictive analysis. We are then clearly moving our problem towards a more synergistic way to code a signal.

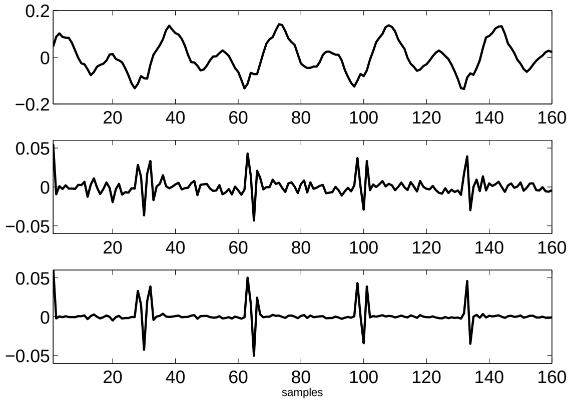

Fig. 2. *Comparison between true 1-norm 10th order LP residual (middle) and iteratively reweighted 1-norm LP residual (bottom) according to Algorithm 1. The original voiced speech is shown on top. Three iteration where performed, sufficient to reach convergence.*

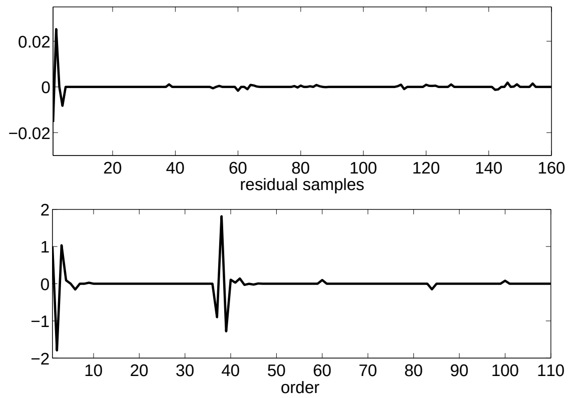

Fig. 3. *Residual and 110th order linear predictor at the convergence of Algorithm 2 after three iterations. The speech segment analyzed is the same as Figure 2.*

6.2. Reweighted Residual and Predictor

The most interesting case is when both predictor and residual are processed in the reweighted minimization. As shown in our previous work [10], the high order predictor is split into the long-term and short-term component through a simple deconvolution. The short-term predictor $A_{stp}(z)$ will have order $N_{stp} = 10$ and the long term predictor (pitch predictor) $P(z) = 1 - g_p z^{-T_p}$ will have order one. The choice of $K = 110$ in (5) means that we can cover accurately pitch delays in the interval $[N_{stp} + 1, K - N_{stp} - 1]$, including the usual range for the pitch frequency $[70\text{Hz}, 500\text{Hz}]$.

In the coding process, we can make a distinction between the voiced case and the unvoiced case. In particular, when the pitch gain g_p is lower than a certain threshold, we will not code the long term

Table 1. Comparison between the MPE residual estimation methods in terms of Segmental SNR and Mean Opinion Score (PESQ evaluation). A 95% confidence intervals is given for each value.

METHOD	SSNR	MOS
MPE1r	20.9±1.9	3.24±0.03
MPE1	20.0±3.2	3.20±0.12
MPE2r	19.3±2.9	3.17±0.10
MPE2	18.5±2.1	3.17±0.22

informations and we will allocate more pulses for the residual, usually less sparse than the voiced residual. In our experimental analysis we have set the threshold to $TH_{g_p} = 0.05$. $M = 5$ and $M = 10$ pulses are used respectively in the voiced and unvoiced case. Just like we did in Section 6.1, the positions of the M pulses of largest magnitude are used in the Analysis-by-Synthesis to define the only nonzero samples. The quantization procedure is also the same as in Section 6.1, except for the quantization of T_p and g_p for which we use respectively 7 and 6 bits. This produces a bit rate of 5450 bit/s in the voiced case and 4900 bit/s in the unvoiced case, and an approximate average bit rate of 5175 bit/s. We will compare our method (**J11r**) with the scheme without the reweighting (**J11**) presented in Equation (5) and the method where the significant coefficients are chosen using a model order selection procedure [3] (**J11os**), we also compared the method with both reweighting and model order selection. In the reweighting cases, only three iteration were needed to reach convergence in all the analyzed frames. The results shown in Table 2, demonstrate a net improvement over the traditional method (**J11**) and a slight improvement also over (**J11os**), without the costly model order selection procedure. The combinations of both methods (**J11r+os**), shows the best results. This is due to the combination of the reweighting procedure that “concentrates” the nonzero parts in the high order polynomial with the model order selection that “spots” the important ones.

Table 2. Comparison between the coding methods with joint estimation of residual and predictor in terms of Segmental SNR and Mean Opinion Score (PESQ evaluation). A 95% confidence intervals is given for each value.

METHOD	SSNR	MOS
J11r+os	27.9±0.9	3.59±0.02
J11r	25.3±1.3	3.43±0.03
J11os	24.7±1.0	3.40±0.09
J11	23.9±1.9	3.22±0.09

7. CONCLUSIONS

In this paper, we have proposed a method to enhance sparsity in linear prediction based on the reweighted 1-norm error minimization. With just few iterations, we were able to move the error minimization criterion toward the 0-norm solution, showing general improvements over conventional 1-norm minimization in coding purposes. Statistical reasons supporting the new criterion have also been provided. A concluding remark would also be that in the cases analyzed, we have no prior knowledge of where the residual should be nonzero. This brings the bit allocated to describe the position of few samples to significantly increase the rate. An interesting case, that would subject to further analysis would be to *structure* the reweighting process by imposing where we would like to have the nonzero pulses located. First experiments have shown to be promising and will be subject of our future work.

8. REFERENCES

[1] J. Makhoul, “Linear Prediction: A Tutorial Review”, *Proc. IEEE*, vol. 63(4), pp. 561–580, Apr. 1975.

[2] D. Giacobello, M. G. Christensen, J. Dahl, S. H. Jensen and M. Moonen, “Sparse Linear Predictors for Speech Processing,” *Proc. Interspeech*, 2008.

[3] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen “Speech Coding Based On Sparse Linear Prediction”, to appear in *Proc. European Signal Processing Conference*, 2009.

[4] B. S. Atal and J. R. Remde, “A new model of LPC excitation for producing natural sounding speech at low bit rates”, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 7, 1982, pp. 614 – 617.

[5] P. Kroon, E. D. F. Deprettere, and R. J. Sluyter, “Regular-pulse excitation - a novel approach to effective multipulse coding of speech”, *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34, pp. 1054–1063, 1986.

[6] M. R. Schroeder and B. S. Atal,“Code-excited linear prediction (CELP): high-quality speech at very low bit rates,” *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 10, pp. 937940, 1985.

[7] E. J. Candés, M. B. Wakin, and S. P. Boyd, “Enhancing sparsity by reweighted ℓ_1 minimization,” *Journal of Fourier Analysis and Applications*, vol. 14(5-6), pp. 877–905, 2008.

[8] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[9] P. Stoica and R. Moses, *Spectral Analysis of Signals*, Pearson Prentice Hall, 2005.

[10] D. Giacobello, M. G. Christensen, J. Dahl, S. H. Jensen, and M. Moonen, “Joint estimation of short-term and long-term predictors in speech coders”, in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 2009.

[11] P. C. Hansen and D. P. O’Leary, “The use of the L-curve in the regularization of discrete ill-posed problems”, *SIAM Journal on Scientific Computing*, vol. 14, no. 6, pp. 1487–1503, 1993.

[12] S. Gazor and W. Zhang, “Speech probability distribution”, *IEEE Signal Processing Letters*, vol. 10, no. 7, pp. 204–207, 2003.

[13] J. Lansford and R. Yarlagadda, “Adaptive L_p approach to speech coding,” in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 1, pp. 335–338, 1988.

[14] A. D. Subramaniam, B. D. Rao, “PDF optimized parametric vector quantization of speech line spectral frequencies”, *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 2, 2003.