A SPEECH SPECTRUM DISTORTION MEASURE WITH INTERFRAME MEMORY

Fredrik Nordén and Thomas Eriksson

Chalmers University of Technology
Department of Signals and Systems
412 96 Göteborg, Sweden

ABSTRACT

In this paper we present a novel spectral distortion measure with interframe memory. The memory gives the possibility to take into account the dynamics of the time evolution of the speech spectrum, which has shown to have a significant importance on the perceived speech quality.

Memory is introduced by linear filtering of the time evolution of the difference log spectrum. This facilitates smoothing of spectrum with a kept ability to track quick transitions.

Our results point at a substantially improved performance when rapidly evolving spectrum errors are punished in the measure.

1. INTRODUCTION

In most modern speech coding systems the parameters from linear prediction (LP) analysis are used to represent the spectrum of the speech signal. The LP coefficients are typically calculated on a frame by frame basis, usually updated every 20 ms, using a 10-12th order linear prediction analysis. A large share of the total bitrate in low bit rate speech coding systems is used for transmitting the LP parameters. Thus, efficient coding of these parameters is important, and much work has been devoted to this area [1, 2, 3].

Encoding of spectrum introduces distortion. In order to evaluate this distortion, several distortion measures have been proposed. The prevalent distortion measure is the *spectral distortion measure* (SD),

$$\mathrm{S}D_n^2 = \frac{20^2}{2\pi}\int_{-\pi}^{\pi}\left(\log_{10}\left|H_n\left(\omega\right)\right| - \log_{10}\left|\tilde{H}_n\left(\omega\right)\right|\right)^2 d\omega, \tag{1}$$

where $H_n(\omega)$ and $\tilde{H}_n(\omega)$ are the original and the coded spectrum for frame n, respectively. When evaluating a coder, SD is calculated for each frame and then averaged over the frames in the evaluation set. In [1], Paliwal and Atal stated that an average SD of around 1 dB is required for transparent spectrum coding quality.

Even though the SD measure is widespread it has some drawbacks. It processes each frame independently and therefore fails to take the time evolution of the spectrum parameters into account. Therefore two different coding methods may not be distinguished comparing SD numbers, but still attain different perceptual ratings.

The importance of the time evolution of spectrum parameters has been the motivation for a number of studies on speech spectrum quantization. Knagenhjelm and Kleijn [4] proposed a method that counteracts the increased fluctuations in the spectrum trajectories due to quantization. Their method smoothes the spectrum trajectories in the decoder with the constraint that the decoded spectrum vector should remain in the same Voronoi region as the received vector. Another approach was taken by Kleijn and Hagen [5] who proposed a method incorporating the dynamics of the trajectories in the encoder operation, i.e. constrained search. They obtained a perceptual improvement by reducing the spectral distance of successive frames at the cost of an increased average spectral distortion. In [6], Samuelsson *et al.* compared the above mentioned methods. Both methods were found to enhance the perceptual performance of the coder.

The previously proposed methods are vector quantization (VQ) methods, and are evaluated in terms of their performance in a VQ framework. As opposed to the previous methods, the goal with this report is not to propose another VQ method, but instead to propose and evaluate a simple parameter- and quantizer-independent quality measure, well suited for performance analysis, but also suitable to derive new VQ methods from [7][1]. We have chosen to incorporate memory in the well-known and wide-spread SD measure (1), by including linear filtering of the log spectrum difference.

The novel measure is defined in section 2 and in section 3, we discuss properties of the proposed measure as compared to previously proposed methods. In section 4, the perceptual performance of the new measure is determined by listening tests, and the conclusions are given in section 5.

[1] In [7] two VQ methods inspired by this novel measure were presented.

2. A MEMORY BASED SD MEASURE

We propose a generalized SD measure (1) with interframe memory, *Spectral Distortion with interframe Memory* (SDM), that facilitates control of the speech spectrum evolution, by controlling the evolution of the spectrum error, c.f. eq. 1,

$$e_n(\omega) = 20\log_{10}|H_n(\omega)| - 20\log_{10}|\tilde{H}_n(\omega)|. \quad (2)$$

The SDM measure can be seen as a two step system. First order FIR filtering of the process $e_n(\omega)$ followed by an evaluation of the SD integral give us

$$\mathrm{SDM}_n^2 = \frac{1}{2\pi(1+b^2)}\int_{-\pi}^{\pi}\left(e_n(\omega) - be_{n-1}(\omega)\right)^2 d\omega, \quad (3)$$

where b is a parameter to control the spectrum error interframe correlation and the filter gain is normalized to unity. This linear filtering facilitates the ability to control the time evolution of the spectrum error, $e_n(\omega)$.

By computing the expected value of SDM_n^2 and write it as a function of the variance of the spectrum error,

$$\sigma_e^2(\omega) = \mathrm{E}[e_n^2(\omega)], \quad (4)$$

and the variance of the differential spectrum error,

$$\sigma_d^2(\omega) = \mathrm{E}[(e_n(\omega) - e_{n-1}(\omega))^2], \quad (5)$$

we get

$$\mathrm{E}[\mathrm{SDM}_n^2] = \frac{1}{2\pi(1+b^2)}\int_{-\pi}^{\pi}\left((1-b)^2\sigma_e^2(\omega) + b\sigma_d^2(\omega)\right)d\omega. \quad (6)$$

From this equation, we see that varying the parameter b between 0 and 1 corresponds to a compromise between minimizing $\sigma_e^2(\omega)$ and $\sigma_d^2(\omega)$. The SD measure is a special case of the SDM measure with $b = 0$, where all focus is on the variance of the spectrum error.

3. PROPERTIES AND COMPARISON

To gain insight about the features of the proposed measure, we discuss it in a quantization scenario. Time dependent distortion, such as increased fluctuations in the spectrum sequence due to quantization, has in several previous studies [4, 5, 7] been shown to be perceptually annoying, maybe even more so than rough quantization. Smoothing of spectrum sequences is therefore a preferred action as long as quick transitions in the original sequence are tracked. A simple low-pass [2] filtering of the distorted spectrum is not a good approach, as it would prevent the ability to track quick transitions in the original spectrum sequence. This has been recognized in e.g. [4, 5], where non-symmetric methods have been utilized to avoid over-smoothing.

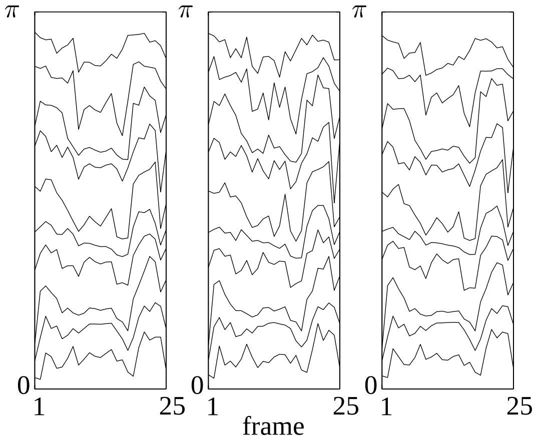

Fig. 1: Line spectral frequency (LSF) trajectories for 0.5 seconds of speech (the word *task*). From left to right: Original spectrum sequence, a spectrum sequence with a non correlated spectrum error sequence ($b = 0$), and a spectrum sequence with a correlated spectrum error sequence ($b = 0.9$). Both the distorted sequences have the same average SD.

As opposed to the solutions above, our approach does not work directly on the spectrum sequence, it works with linear combinations of the spectrum error. Minimization of the SDM measure clearly occurs when $e_n(\omega)$ is selected to be as close to $be_{n-1}(\omega)$ as possible, which leads to a correlated spectrum error sequence, depending on the filter parameter b. In a quantization context, SDM will choose spectrum code vectors such that the quantization noise has an interframe correlation close to b, since such a correlation leads to the minimum SDM value, see section 4.

The SDM measure ($b > 0$) punishes rapid variations in the spectrum error sequence. The effect of this can be described as imposing low-pass characteristics upon the time evolution of the spectrum error $e_n(\omega)$. Basically, the low-pass character of the spectrum error means that the quantized spectrum $\tilde{H}_n(\omega)$ inherits the evolutionary high-pass part directly from the spectrum $H_n(\omega)$, without being affected by $e_n(\omega)$. This is advantageous both during stable voiced sounds, where a slowly evolving spectrum is critical for the perceived quality, and during onsets, where it is important that the spectrum can evolve more freely.

The smoothing property of the SDM measure is visualized in figure 1, where the distorted spectrum sequence with a correlated spectrum error sequence, $b = 0.9$, clearly have smoother trajectories in regions where the original spectrum sequence is smooth. In section 4 listening tests show that

[2] In this context, low-pass means evolutionary low-pass, i.e. the signal has a high correlation from one frame to the next.

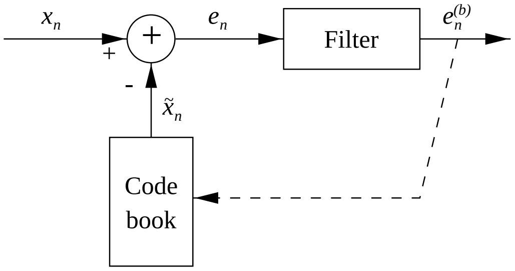

Fig. 2: An example of a quantizer employing the proposed SDM measure in the codeword search. The quantizer output $\tilde{x}_n$ is chosen to minimize the average energy of $e_n^{(b)}$.

given a constant average SD, a correlated error is subjectively preferable.

In the next section, we follow this brief explanation of the properties of the SDM measure, with a subjective evaluation.

4. SUBJECTIVE PERFORMANCE

We have performed a series of listening tests, to evaluate the perceptual performance of the SDM measure (3) with respect to the choice of the parameter b.

In order to avoid a dependency on a specific LP parameterization or quantization scheme, quantization of spectrum vectors was simulated by adding correlated noise in the log spectral domain, c.f. eq. 2,

$$20\log_{10}|\tilde{H}_n(\omega)| = 20\log_{10}|H_n(\omega)| - e_n^{(\rho)}(\omega), \qquad (7)$$

where $e_n^{(\rho)}$ is noise with interframe correlation ρ,

$$e_n^{(\rho)}(\omega) = u_n(\omega) + \rho e_{n-1}^{(\rho)}(\omega), \qquad (8)$$

and $u_n(\omega)$ represents white noise. This setup can be motivated by studying a real quantizer where the codewords are selected to minimize the SDM measure. An example of such a quantization system is given in Figure 2. The quantizer selects codewords to minimize the filtered output

$$e_n^{(b)}(\omega) = e_n(\omega) - be_{n-1}(\omega), \qquad (9)$$

and it is easy to show that for high-rate quantizers [8], the filter output, $e_n^{(b)}$, will be approximately white. This means that the error signal $e_n(\omega)$ (2) will have a correlation close to the correlation produced by the inverse of the filter in the SDM measure:

$$e_n(\omega) = e_n^{(b)}(\omega) + be_{n-1}(\omega), \qquad (10)$$

where $e_n^{(b)}(\omega)$, as previously mentioned, is approximately white in a high-rate quantization scenario.

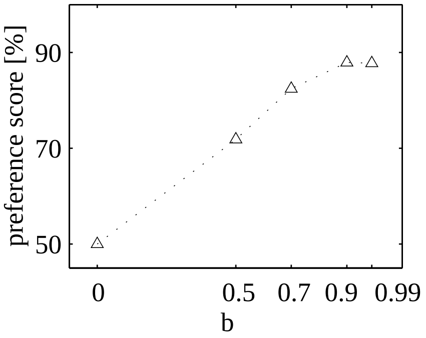

Fig. 3: Subjective evaluation as a function of the filter coefficient, b. The preference scores indicate the preference for noise with positive correlation, compared to uncorrelated noise.

To evaluate different settings of the parameter b in the SDM measure we changed the interframe correlation, ρ, of the added noise to match the correlation giving the minimum value of the SDM measure, i.e. $\rho = b$. The noisy spectral vectors were used to synthesize the speech files used in the listening tests.

We prepared files with eight sentences. Four different parameter settings were compared. A total of 7 listeners made pairwise forced choice comparisons between the different settings using headphones.

4.1. Choice of filter

To illustrate the advantages of incorporating memory in the SD measure, we have performed listening tests with varying values of the parameter b, i.e. varying interframe correlation in the spectrum error, while the average SD was kept constant. Figure 3 clearly shows that even though the SD value is constant, the perceptual quality of the distorted sentences varies substantially with the value of b, and that high correlation between consecutive spectrum errors ($b > 0$), is preferable. A statistical evaluation in the form of a series of t-tests [9] revealed that there were significant preference for $b \neq 0$ at a significance level of 1 %.

4.2. Gains

We have also performed listening tests to estimate the maximum possible gains of the SDM measure compared to the SD measure, i.e SDM with $b = 0$. The noise power for the reference system $b = 0$, was set to a value corresponding to $\mathrm{SD} = 1.33$ dB. This seemingly high distortion was, while clearly audible for $b = 0$, still transparent to the listeners when perceptually correlated noise, $b = 0.9$, were used. The results in figure 4 show that we can increase the average SD with 0.9 dB for the correlated noise case, and

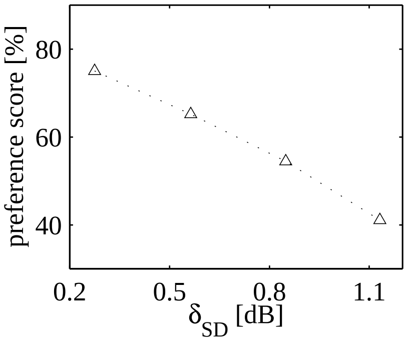

Fig. 4: Subjective evaluation of the maximum gains of using a SDM measure with $b = 0.9$ compared to the SD measure, i.e. $b = 0$. The δ_{SD} number represents the increase in SD for the $b = 0.9$ relative the $b = 0$ measure with SD $= 1.33$ dB.

still get the same subjective quality as for the uncorrelated case.

5. CONCLUSIONS

We have addressed the problem of finding a perceptually tractable distortion criterion for speech spectrum coding. The proposed criterion punishes quick transitions in the evolution of the spectrum error process. This results in smoother spectrum trajectories at steady state with a kept ability to follow quick transitions at onsets. The listening tests point at a maximum gain in SD of 0.9 dB if the proposed distortion criterion is exploited.

6. REFERENCES

[1] K. K. Paliwal and B. S. Atal, “Efficient vector quantization of LPC parameters at 24 bits/frame,” *IEEE Trans. Speech Audio Processing*, vol. 1, no. 1, pp. 3 – 14, 1993.

[2] W. R. Gardner and B. D. Rao, “Theoretical analysis of the high-rate vector quantization of LPC parameters,” *IEEE Trans. Speech Audio Processing*, vol. 3, no. 5, pp. 367–381, 1995.

[3] A. H. Gray and J. D. Markel, “Distance measures for speech processing,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 24, no. 5, pp. 380–391, 1976.

[4] H. P. Knagenhjelm and W. B. Kleijn, “Spectral dynamics is more important than spectral distorsion,” in *Proc. International Conference on Acoustics, Speech, and Signal Processing*, 1995, vol. 1, pp. 732 – 735.

[5] W. B. Kleijn and R. Hagen, “On memoryless quantization in speech coding,” *IEEE Signal Processing Letters*, vol. 3, no. 8, pp. 228 – 230, 1996.

[6] J. Samuelsson, J. Skoglund, and J. Lindén, “Controlling spectral dynamics in LPC quantization for perceptual enhancement,” in *Proc. 31st Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, 1997.

[7] F. Nordén and T. Eriksson, “Perceptual spectrum quantization,” in *Proc. 34th Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, 2000.

[8] Robert M. Gray, *Source Coding Theory*, Kluwer Academic Publishers, 1990.

[9] J. A. Rice, *Mathematical Statistics and Data Analysis*, Wadsworth & Brooks/Cole Advanced Books & Software, 1988.