ISCAS 2000 - IEEE International Symposium on Circuits and Systems, May 28-31, 2000, Geneva, Switzerland

SUBBAND STATIONARITY ANALYSIS OF SPEECH SIGNALS

R. da S. Maia, F. G. V. Resende, Jr., and S. L. Netto

Programa de Engenharia Elétrica/COPPE, DEL/EE
Universidade Federal do Rio de Janeiro
PO Box 68504, Rio de Janeiro, RJ, 21945-970, Brazil
ranniery@lps.ufrj.br, gil@lps.ufrj.br, sergioln@lps.ufrj.br

ABSTRACT

A subband analysis of the stationarity characteristics of speech signals is performed. The analysis is based on the evaluation of seven distance measure techniques between consecutive speech segments. The overall ensemble comprised a total of 600 speech sentences each of duration varying from 2 s to 3 s, generated by two male and one female speakers. Segment lengths of 10 ms to 30 ms were considered. The experiments have shown that: most distance measure techniques yielded equivalent results; segments of 10 ms presented greater level of stationary between 0 and 1 kHz; for other lengths, all bands presented similar stationarity degrees. Results can be applied for a proper subband speech processing (e.g., coding) that depends on the stationary assumption of the signals involved.

1. INTRODUCTION

Speech coding has become an even more intense research area in the past two decades. Before that, vocoders (codecs for speech signals) attained a high level of compression at the expense of a considerably low quality of the decoded signal. In 1985, Schroeder and Atal [1] changed such status quo by introducing the concept of code-excited linear prediction (CELP) vocoders that achieved a high level of compression rate with a speech quality comparable to standard wave-format coders, such as PCM and its variations. Recently, lower and lower rates are achieved by the speech oders being used in the areas of mobile phones and computer networks [2]. Such enhanced coding efficiency, with respect to the overall distortion rate and computational complexity, is being achieved by better exploring the speech signals intrinsic (statistical) characteristics. This paper focuses on subband analysis of the stationarity characteristics of speech with applications to subband coding.

Seven distance measure techniques found in the literature are used. An ensemble of speech samples is employed, comprising a total of 600 sentences from three different speakers, two male and one female. Several segment lengths are evaluated, ranging from 10 ms to 30 ms. It is verified that most figures of merit yield very similar results. In addition, the subband stationarity assessment is shown to be essencially speaker independent for a given gender. It is also verified that for most segment lengths considered no level of stationarity distinction can be observed between separate frequency bands. However, for segments of 10 ms, it is clearly noticed that the lowpass band, between 0 and 1 kHz, presents a higher level of stationarity than the other bands.

This work was partially supported by FUJB-UFRJ/Brazil.

Organization of the present work is as follows. In Section 2, we describe the subband analysis procedure. In Section 3, the different distance measure techniques used in this work are presented. In Section 4, all experiments performed are described and the corresponding results are shown. Section 5 emphasizes the important conclusions drawn from the analysis provided.

2. SUBBAND STATIONARITY ANALYSIS

The procedure for performing stationarity analysis on the subbands of speech signals is based on measuring some form of variation in the subbands of speech signals. Such variation can be determined, for instance, by the differences between the envelopes of the magnitude responses of each frequency band in consecutive speech frames.

We then first perform a signal decomposition using an analysis filter bank that generates the signal components $s_i(n)$, for $i = 1, \ldots, N$, where N is the total number of subband filters. Following, all signals $s_i(n)$ are segmented into frames of constant length. For each individual frame, a 16-th order linear prediction analysis is performed [3], based on the autocorrelation method using the Hamming window function centered around the segment. The magnitude response of the linear prediction model corresponds to a smoothed version of the speech magnitude spectrum [3],[4]. We then determine for each band, the distances $\mathbf{d_i}$ between these spectral envelopes for all consecutive frames. For a total of K frames, these vectors $\mathbf{d_i}$ have $K-1$ entries. A figure of merit, the so-called degree of spectral variation (DSV), based on such distance measurement is the average value of the elements in $\mathbf{d_i}$, that is

$$DSV_i = \frac{1}{K-1} \sum_{k=1}^{K-1} d_i(k), \tag{1}$$

where $d_i(k)$ represents the k-th entry of $\mathbf{d_i}$.

For a given subband i, the relative variation percentage RVP_i with respect to all N bands is defined as

$$RVP_i(\%) = \frac{DSV_i}{\sum_{i=1}^{N} DSV_i} \times 100 \tag{2}$$

Such figure represents a more meaningful measurement for all comparison purposes.

In this particular paper, two kinds of filter filter bank were employed for speech decomposition. One of them was a 4-band cosine-modulated uniform filter bank derived from a 32-length prototype filter as given in [5]. The other one was a 4-band filter

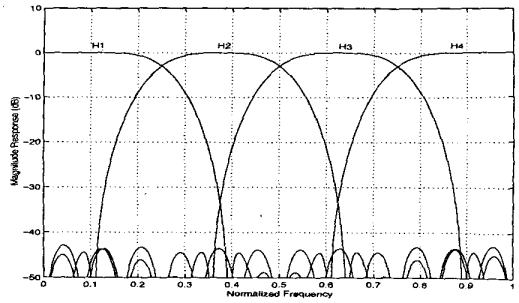

Figure 1: Magnitude response of the 4-band cosine-modulated uniform filter bank used in the decomposition.

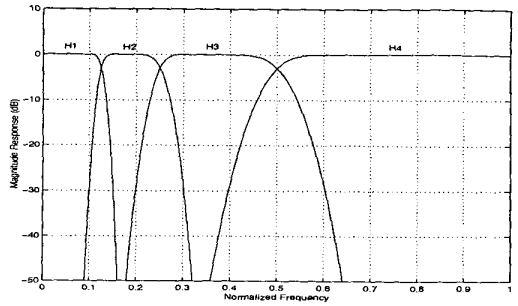

Figure 2: Magnitude response of the 4-band filter bank based on the extremal-phase 90-th order Daubechies wavelet used in the decomposition.

bank based on the extremal-phase 90-th order Daubechies wavelet [5]. These banks are power-complementary such that

$$\sum_{i=1}^{N} |H_i(e^{j\theta})|^2 = 1, \ \forall\theta \in [0, \pi] \tag{3}$$

where $|H_i(e^{j\theta})|$ represents the magnitude response of the i-th filter in the bank. Fig. 1 and Fig. 2 depict the magnitude responses of both filter banks here employed.

3. DISTANCE MEASURE TECHNIQUES

The distance measures determine the difference between some statistical characteristic of consecutive speech frames.

3.1. Log-Spectral Distances

Given two models $\sigma/A(z)$ and $\sigma'/A'(z)$ of the form

$$A(z) = 1 - \sum_{i=1}^{P} a_i z^{-i}; \ A'(z) = 1 - \sum_{i=1}^{P} a'_i z^{-i} \tag{4}$$

where P is the model order and the a_i and a'_i are the corresponding linear prediction coefficients and σ, σ' are gain factors. The error or difference between these two models in a log-magnitude scale is given by

$$V(\theta) = \ln\left[\frac{\sigma^2}{|A(e^{j\theta})|^2}\right] - \ln\left[\frac{(\sigma')^2}{|A'(e^{j\theta})|^2}\right] \tag{5}$$

In [6], a proper distance measurement D_p based on this error function is defined as

$$(D_p)^p = \frac{1}{2\pi}\int_{-\pi}^{\pi} |V(\theta)|^p d\theta \tag{6}$$

For $p = 1$, one has the absolute log-spectral measurement; for $p = 2$; one has the mean-squared log-spectral measurement; and for $p \to \infty$, one has the maximum log-spectral measurement. In [6] it is verified that the distances D_2 and D_∞ are very similar for all practical purposes. In this paper, we use D_2 in dB, given by

$$D_2(dB) = \frac{10}{\ln 10}\sqrt{\frac{1}{L}\sum_{k=-L/2}^{L/2}\left\{\ln\left|\frac{\sigma}{A(e^{\frac{jk\pi}{L}})}\right|^2 - \ln\left|\frac{\sigma'}{A'(e^{\frac{jk\pi}{L}})}\right|^2\right\}^2} \tag{7}$$

where L is the number of points for evaluating (6).

If the two gain factors σ and σ' are normalized by the frame energy, a slightly different measurement results:

$$D'_2(dB) = \frac{10}{\ln 10}\sqrt{\frac{1}{L}\sum_{k=-L/2}^{L/2}\left\{\ln\left|\frac{\tau}{A(e^{\frac{jk\pi}{L}})}\right|^2 - \ln\left|\frac{\tau'}{A'(e^{\frac{jk\pi}{L}})}\right|^2\right\}^2} \tag{8}$$

with

$$\tau = \frac{\sigma}{e}; \ \tau' = \frac{\sigma'}{e'} \tag{9}$$

where e and e' are the energy values of the frames associated to $\sigma/A(z)$ and $\sigma'/A'(z)$, respectively, that is

$$e = \sum_{n=0}^{N-1} s^2(n); \ e' = \sum_{n=0}^{N-1} s'^2(n) \tag{10}$$

3.2. Cepstral Distances

The cepstral coefficients c_i can be obtained from the linear prediction coefficients a_i using [6]:

$$c_i = \begin{cases} a_i + \sum_{k=1}^{i-1} \frac{k}{i} c_k a_{i-k}, & \text{for } 1 \le i \le P \\ \sum_{k=1}^{p} \left(1 - \frac{k}{i}\right) c_{i-k} a_k, & \text{for } P < i \le P' \end{cases} \tag{11}$$

When the equation above is employed, the cepstral coefficients are referred-to as linear-prediction cepstral coefficients. A spectral distance between two sets of cepstral coefficients, $\mathbf{c}$ and $\hat{\mathbf{c}}$, corresponding to consecutive speech frames can then be defined as [6]

$$D_C(dB) = \frac{10}{\ln 10}\sqrt{(c_0 - \hat{c}_0)^2 + 2\sum_{i=1}^{P'} [c_i - \hat{c}_i]^2} \tag{12}$$

where

$$c_0 = \ln\left\{\sigma^2\right\};\ \hat{c}_0 = \ln\left\{\sigma'^2\right\} \tag{13}$$

A simple variation of such measure is given by [7]

$$D_C'(dB) = \frac{10}{\ln 10}\sqrt{2\sum_{i=1}^{P'}[c_i - \hat{c}_i]^2} \tag{14}$$

Clearly, $D_C(dB) = D_C'(dB)$, when $\sigma = \sigma'$.

It can be verified that the cepstral distance in (12) is used as a simpler alternative form of calculating $D_2(dB)$ [6].

3.3. Itakura Distance

The Itakura distance [8] is widely used to determine the distance between two sets of linear prediction coefficients. It is given by [3]

$$D_I(dB) = 10\log\frac{\hat{\alpha}^{\mathbf{T}}\mathbf{R}\hat{\alpha}}{\alpha^{\mathbf{T}}\mathbf{R}\alpha} \tag{15}$$

where $\mathbf{R}$ is the autocorrelation matrix of the corresponding speech frame, and α and $\hat{\alpha}$ are derived from the vectors $\mathbf{a}$ and $\hat{\mathbf{a}}$, containing the linear prediction coefficients of the true model and of the alternative model, respectively, as

$$\alpha = \begin{bmatrix} 1 \\ -\mathbf{a} \end{bmatrix};\ \hat{\alpha} = \begin{bmatrix} 1 \\ -\hat{\mathbf{a}} \end{bmatrix} \tag{16}$$

In this paper, the true and the alternative models corresponded to consecutive speech frames. It can be observed that D_I is not symmetric as $D_I(\mathbf{a}, \hat{\mathbf{a}}) \neq \mathbf{D_I}(\hat{\mathbf{a}}, \mathbf{a})$. To overcome such drawback, a new measure can be defined as the average between these two distances, that is [3]

$$D_I' = \frac{1}{2}\left[D_I(\mathbf{a}, \hat{\mathbf{a}}) + \mathbf{D_I}(\hat{\mathbf{a}}, \mathbf{a})\right] \tag{17}$$

The Itakura Distance is based in the ratio between energies of residual signals, resulting from filtering the speech frame $s(n)$ by the true and alternative models $1/A(z)$ and $1/A'(z)$, respectively. Relationship between D_I and D_2 is highly complex and essentially nonlinear as shown in [6].

3.4. Itakura-Saito Distance

The Itakura-Saito distance is a simplification of the Itakura distance given in [8], resulting in

$$D_{IS}(dB) = 10\log\frac{(\alpha_o - \hat{\alpha})^{\mathbf{T}}\mathbf{R}(\alpha_o - \hat{\alpha})}{\alpha_o{}^{\mathbf{T}}\mathbf{R}\alpha_o} \tag{18}$$

Such expression results from a hypothesis test where it is assumed that the linear prediction coefficients are jointly-normally distributed with mean given by true linear prediction vector $\mathbf{a_o}$. It is then assumed that $\mathbf{a} \approx \mathbf{a_o}$, what is not essentially true due to the stochastic nature of the linear prediction process [3].

As in the case of D_I, the Itakura-Saito distance is not symmetric. Once again, this can be overcome by defining an average distance measurement as

$$D_{IS}' = \frac{1}{2}\left[D_{IS}(\mathbf{a}, \hat{\mathbf{a}}) + \mathbf{D_{IS}}(\hat{\mathbf{a}}, \mathbf{a})\right] \tag{19}$$

3.5. LAR Distance

Due to quantization, linear prediction coefficients are often transformed into reflection coefficients, $\{k_i : i = 1, \ldots, P\}$, in coding applications. In practice, the reflection coefficients arise naturally by using the Levinson-Durbin algorithm to solve the linear prediction problem, when the autocorrelation method is employed. Such reflection coefficients are also commonly transformed in practice into a new set of coefficients, the so-called log-area ratio (LAR) coefficients, through the relationship [4]

$$l_i = \log\left[\frac{1 + k_i}{1 - k_i}\right] \tag{20}$$

The distance between two sets of LAR coefficients is commonly defined as the Euclidean norm of the difference between them, that is

$$D_{LAR}(dB) = 10\log\left[\sqrt{\frac{1}{P}\sum_{i=1}^{P}(l_i - \hat{l}_i)^2}\right] \tag{21}$$

In [3], it is mentioned that among all objective measurements of speech quality, the LAR distance is the most correlated to subjective measurements.

4. SUBBAND ANALYSIS

All subsequent analyses were performed based on three sets of 200 short sentences, varying in length from 2 to 3 s, phonetically balanced for the Brazilian Portuguese of Rio de Janeiro [9]. Two sets are from male speakers and one is from a female speaker.

The values of all distances were determined. Namely: D_2 (7), D_2' (8) with $L = 512$; D_C (12) and D_C' (14), with $P' = 32$; D_I' (17) and D_{IS}' (19); and D_{LAR} (21). All linear prediction analyses used 16 coefficients, and were based on the autocorrelation method with the Hamming window [3].

Table 1 shows the values of RVP, defined in (2), for each distance measure technique for each subband. These values were obtained for 200 sentences of a single male speaker, and the segment length was made constant at 10 ms. The sample frequency was 8 kHz and it was used the 4-band cosine-modulated filter bank, thus determining the four subbands in the ranges 0-1 kHz, 1-2 kHz, 2-3 kHz, and 3-4 kHz, respectively. Analysis of Table 1 suggests that the outcomes obtained through different distance measure methods are almost similar, and that the degree of stationarity of speech signals is greater in the frequency range 0-1 kHz.

Table 2 shows the RVP's for the D_2 considering a total of 600 sentences for three different speakers, 200 sentences each, where two, M1 and M2, were of male gender and one of female gender, F1. The segment lengths and the analysis bank are the same as those used for Table 1. From Table 2 it can be observed that for the condition employed the degree of stationarity is higher in the lowpass band.

Table 3 and Table 4 shows the RVP's for the D_2 measurement resulting from tests with 600 sentences of all three speakers (200 sentences each) for several frame lengths, ranging from 10 ms to 30 ms, using the two filter banks described previously. For the wavelet filter bank, the four subbands BW1, BW2, BW3, and BW4, were in the ranges 0-500 Hz, 500-1000 Hz, 1-2 kHz, and 2-4 kHz, respectively. Table 3 shows that for segments of 10 ms, the BW1 band has a distinctive greater level of starionarity. But for

Table 1: RVP for each distance measure technique for 200 of a single male speaker for speech segments of 10 ms, when the cosine-modulated filter bank is used.

	BW1	BW2	BW3	BW4
D_2	29.26	24.54	23.13	23.07
D_2'	28.32	24.62	23.22	23.84
D_C	24.91	26.10	24.66	24.33
D_C'	29.41	24.53	22.97	23.09
D_I'	26.28	26.06	24.37	23.33
D_{IS}'	29.30	22.36	23.40	24.94
D_{LAR}	23.43	27.26	26.58	22.73

Table 2: RVP for 200 sentences of each speaker for speech segments of 10 ms, when the cosine-modulated filter bank is used.

	BW1	BW2	BW3	BW4
M1	29.26	24.54	23.13	23.07
M2	29.14	24.00	22.49	24.37
F1	27.12	26.04	23.37	23.47

others speech segments the bands are approximately equally stationary. Table 4 shows that when the wavelet filter bank is used, there are no significant stationarity differences among the frequency bands.

5. CONCLUSION

In this work, a thorough subband analysis of the stationary characteristics of speech signals was performed. Several distance measure techniques were used and a speech ensemble comprising 600 sentences from 2 s to 3 s each one was employed. Results have demonstrated the folowing: most of distance measures yielded equivalent results. It was verified that for segments of 10 ms the lowpass band presented a higher level of stationarity when a cosine-modulated uniform filter bank is used. When a extremal-phase 90-th order Daubechies wavelet filter bank is used there are no apparent stationarity differences among the frequencies ranges for all segment lengths. Such results can be applied to computationally-efficient subband coding schemes for speech signals.

6. REFERENCES

[1] 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*, pp. 937–940, Tampa, FL, 1985.

[2] R. Salami, *et al.*, "ITU-T G.729 Annex A: Reduced complexity 8 kb/s CS-ACELP codec for digital simultaneous voice and data," *IEEE Communications*, vol. 35, no. 9, pp. 56–63, Sept. 1997.

[3] J. Deller, J. Proakis, and J. Hansen, *Discrete-time Processing of Speech Signals*, New York:NY, MacMillan, 1993.

Table 3: RVP for 600 sentences for each segment length when the cosine-modulated filter bank is used.

	BW1	BW2	BW3	BW4
10 ms	28.50	24.86	23.00	23.64
15 ms	26.69	25.05	24.02	24.24
20 ms	26.06	25.10	24.41	24.43
25 ms	25.65	25.12	24.64	24.59
30 ms	25.23	25.19	24.87	24.71

Table 4: RVP for 600 sentences for each segment length when the Daubechies wavelet filter bank is used.

	BW1	BW2	BW3	BW4
10 ms	24.98	25.32	25.22	24.48
15 ms	24.31	25.25	25.40	25.04
20 ms	24.11	25.37	25.43	25.09
25 ms	24.14	25.39	25.32	25.15
30 ms	23.85	25.48	25.40	25.27

[4] A. S. Spanias, "Speech coding: a tutorial review," *Proc. IEEE*, vol. 82, pp. 1541–1582, Oct. 1994.

[5] G. Strang and T. Nguyen, *Wavelets and Filter Banks*, Wellesley, MA: Cambridge Press, 1996.

[6] A. H. Gray and J. D. Markel, "Distance measures for speech processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, pp. 380–391, Oct. 1976.

[7] N. Kitawaki and H Nagabuchi, "Objective quality evaluation for low-bit-rate speech coding systems," *IEEE Jrnl. Selected Areas Communications*, vol. 6, no. 2, pp. 242–248, Feb. 1988.

[8] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 23, pp. 67–72, Feb. 1975.

[9] A. Alcaim, J. A. Solewicz, and J. A. de Moraes, "Phone relative frequency and lists of phonetically balanced sentences for the Portuguese language in Rio de Janeiro," *Revista Soc. Brasil. Telecom.* (in Portuguese), vol. 7, no. 1, Dec. 1992.