Modeling of the analytic spectrum for speech recognition.

CONFERENCE PAPER · JANUARY 1989

Source: DBLP

3 AUTHORS:

Climent Nadeu
Polytechnic University of Catalonia
171 PUBLICATIONS **1,150** CITATIONS
SEE PROFILE

Eduardo Lleida
University of Zaragoza
199 PUBLICATIONS **843** CITATIONS
SEE PROFILE

Javier Hernando
Polytechnic University of Catalonia
176 PUBLICATIONS **974** CITATIONS
SEE PROFILE

MODELING OF THE ANALYTIC SPECTRUM FOR SPEECH RECOGNITION

Climent Nadeu, Eduardo Lleida and Fco. Javier Hernando

Dept. Teoria del Senyal i Comunicacions
Universitat Politècnica de Catalunya, Barcelona, Spain

ABSTRACT

In this paper, a new spectral representation is introduced and applied to speech recognition. As the widely used LPC autocorrelation technique, it arises from an optimization approach that starts from a set of M+1 autocorrelations estimated from the signal samples. This new technique models the analytic spectrum (Fourier's transform of the causal autocorrelation sequence) by assuming that its cepstral coefficients are zero beyond M, and uses an extremely simple algorithm to compute the non-zero coefficients. In speech recognition, the same Euclidean cepstral distance measure that is the object of the optimization is also used to calculate the spectral dissimilarity. Preliminary recognition tests with this technique are presented.

INTRODUCTION

Most speech processing systems use some sort of spectral estimation to characterize every signal frame by a small set of parameters [1]. A spectral estimation technique widely used in speech recognition and other signal processing applications is the linear predictive coding (LPC). This technique assumes an all-pole model of the spectrum which arises from a mean square minimization of the linear prediction error of the speech signal [2].

However, there is an alternative way of deriving the all-pole spectral model of the LPC method: the known Burg's optimization approach [3]. According to it, a set of autocorrelations r_n, n = 0, 1, ... , M, are assumed accurate enough and are the constraints of an optimization problem whose objective is to maximize the entropy of the underlying process to which belongs the signal. Consequently, it has been called the maximum entropy method (MEM). Furthermore, the entropy is a flatness measure in the sense that it measures the closeness between the spectrum estimate and a constant (flat) spectrum of the same area (r_0).

This work was supported by the PRONTIC grant number 105/88.

In fact, maximum flatness is a sensible goal for speech analysis because a maximally flat spectrum consistent with the given autocorrelations r_n aims at including only the information carried by these constraints r_n, without adding spurious spectral components. However, there are other ways of measuring flatness different from the entropy measure used by the LPC or ME method [4]. For example, it exists a spectral estimation method [5] which uses a different measure of entropy and provides flatter spectra at peaks than the LPC method; it estimates the spectrum by means of a cepstral smoothing with a data adaptive window of length 2M+1 that preserves the autocorrelations from 0 to M. Unfortunately, this method requires an iterative algorithm to find the spectrum.

Both ME methods, as all the other spectral estimators that arise from the optimization approach, are characterized by the spectral measure which is the object of the optimization. In general, it can be expressed as a functional that measures the distance between two spectra, namely

$$D(S,S_p) = \frac{1}{2\pi} \int_{-\pi}^{\pi} d[S(\omega),S_p(\omega)]\, d\omega \qquad (1)$$

where $S(\omega)$ is the spectrum to be estimated, $S_p(\omega)$ is a known spectrum containing a-priori information about the spectral features of the random process, and $d(x,y)$ is a distance measure between points x and y that determines the way in which the two types of spectral information, r_n and $S_p(\omega)$, are articulated to give an estimate of $S(\omega)$. Obviously, with this approach, $D(S,S_0)$ has to be minimized since it is a measure of error or distance.

When an a-priori spectrum is not known, it seems reasonable to assume $S_0(\omega) = r_0$ (average value of $S(\omega)$), so that (1) turns out to be a measure of separation from the flat spectrum. In this manner, minimizing $D(S,r_0)$ is equivalent to maximize flatness according to the specific form of the distance $d(x,y)$.

On the other hand, a very common problem in speech recognition and coding (vector quantization, speech quality assessment) is to measure the distance existing between the spectra of two speech frames [6]. Thus, a completely coherent approach results from using the same type of distance for both 1) estimating the spectra S_1 and S_2 (with S_0 = constant) separately, and 2) evaluating with (1) the distance between them.

In this paper, we will use a spectral function instead of the spectrum itself. This function is the so-called analytic spectrum [7] which has the same poles as the spectrum. Using as d(x,y) the Euclidean distance, a very simple model based on the cepstrum of this function is obtained and the computational load of the corresponding algorithm is lower than that of the LPC technique. Finally, the results of a preliminary speech recognition test will be given.

THE ANALYTIC SPECTRUM

Let $R^+(n)$ denote the "causal part" of the autocorrelation R(n), namely

$$R^+(n) = 2R(n) \quad \text{for} \quad n > 0 \qquad (2.a)$$
$$= R(n) \quad \text{for} \quad n = 0 \qquad (2.b)$$
$$= 0 \quad \text{for} \quad n < 0 \qquad (2.c)$$

Its Fourier transform $S^+(\omega)$ is

$$S^+(\omega) = S(\omega) + jH(\omega) = |S^+(\omega)|e^{j\theta(\omega)} \qquad (3)$$

where $H(\omega)$ denotes the Hilbert transform of $S(\omega)$. $S^+(\omega)$ may be referred to as the "analytic spectrum" [7], in correspondence with the analogous definition used in amplitude modulation. Accordingly, we may define the spectral envelope

$$E(\omega)=|S^+(\omega)| \qquad (4.a)$$

which has already been used for spectral estimation [7], and the spectral "instantaneous frequency"

$$\tau(\omega) = -\frac{d\theta(\omega)}{d\omega} \qquad (4.b)$$

Since $R^+(n)$ is a minimum phase function, either $E(\omega)$ or $\tau(\omega)$ specify $S(\omega)$ uniquely and viceversa. Moreover, the poles of $S(\omega)$ appear in both $\tau(\omega)$ and the square envelope $E^2(\omega)$.

The complex cepstrum $C^+(n)$ of $R^+(n)$, i.e. the Fourier's series coefficients of $\log S^+(\omega)$, directly characterizes both functions $\log E(\omega)$ and $\tau(\omega)$, since

$$\log E(\omega) = C^+(0) + \sum_{n=1}^{\infty} C^+(n)\cos n\omega \qquad (5)$$

$$\tau(w) = \sum_{n=1}^{\infty} n\, C^+(n)\cos n\omega \qquad (6)$$

On the other hand, due to the causality and minimum phase properties of $R^+(n)$, its cepstrum $C^+(n)$ can be obtained by means of the following recursion [8]

$$C^+(n) = \frac{2}{R(0)}\left[R(n) - \sum_{k=1}^{n-1}\frac{k}{n}C^+(k)R(n-k)\right], \quad n>0 \qquad (7.a)$$

$$= \log R(n), \quad n=0 \qquad (7.b)$$

$$= 0, \quad n<0 \qquad (7.c)$$

where the relationship (2) between $R^+(n)$ and R(n) has been used.

Observe that to compute $C^+(n_1)$, apart from its previous values, only the autocorrelations from lag 0 to lag n_1 are needed. Hence, the first N values of $C^+(n)$ are completely determined by the first N values of R(n) and vice versa. From (7), we find that the reversed relation is

$$R(n) = \frac{1}{2}R(0)C^+(n) + \sum_{k=1}^{n-1}\frac{k}{n}C^+(k)R(n-k), \quad n>0 \qquad (8.a)$$

$$= \exp C^+(n), \quad n=0 \qquad (8.b)$$

$$= R(-n), \quad n<0 \qquad (8.c)$$

MODELING OF THE ANALYTIC SPECTRUM

The analytic spectrum $S^+(\omega)$ and its associated functions may be useful in speech analysis and recognition. Given a spectrum, we can compute and use the spectral functions S^+, E or τ derived from it, or else we can directly find an estimate of them. The last approach is used in a recent paper [9] where the MEM is applied to the envelope $E(\omega)$ instead of the spectrum itself. This means that the constraints of the optimization problem are M+1 Fourier's series coefficients of $E(\omega)$, which are computed from a large set of autocorrelations. The result is a robust spectral representation technique that achieves an equivalent SNR improvement of 13dB in the presented experiments of speech recognition. However, the computational complexity is higher than that of the LPC technique.

Henceforth, we will only consider as starting data a small set of autocorrelations r_n. However, according to (7) and (8), there exists an one to one correspondence between these first M+1 autocorrelation values and the corresponding M+1 first cepstral coefficients of the analytic spectrum. Therefore, we can compute these coefficients C^+_n with (7) and use them as the constraints of the optimization problem. This problem then consists of

$$\text{minimizing} \quad D(T,t_0)$$
$$\text{subject to } C^+(n) = C^+_n, \quad n = 0, \ldots, M$$

where $T(\omega)$ may be $S^+(\omega)$, $E(\omega)$, $\tau(\omega)$, $S(\omega)$ or functions of them and t_0 is the average value of $T(\omega)$. Obviously, each distance measure $d(T,t_0)$ produces a different spectral estimate for the same data. Moreover, when this spectral estimation technique is used in speech recognition, it seems reasonable to employ the same type of distance measure $D(T_1,T_2)$ to calculate spectral distances between frames. Then, the whole process shows a complete coherence.

Although many kinds of spectral models and distances could be considered, we will concentrate in this paper on one of them. We select the spectral function

$$T(\omega) = \frac{d}{d\omega}[\log S^+(\omega)] \qquad (9)$$

and the Euclidean distance

$$d(T_1,T_2) = \left| \frac{d}{d\omega} \log S_1^+(\omega) - \frac{d}{d\omega} \log S_2^+(\omega) \right|^2 \qquad (10)$$

which is perceptually meaningful when used with $S(\omega)$ [10].

Observe that the corresponding global distance (1) takes the form

$$D(T_1,T_2) = \sum_{n=1}^{\infty} [nC_1^+(n) - nC_2^+(n)]^2 \qquad (11)$$

Thus, to find the solution of the minimization problem (8), we can equal to zero the derivatives of $D(T,t_0)$ in (11), with respect to the unknown cepstral coefficients of the analytic spectrum, resulting

$$C^+(n) = 0, \quad n > M \qquad (12)$$

that is, a zero extrapolation of the data beyond M.

In other words, this method of spectral estimation performs a cepstral smoothing, analogously to the alternative MEM mentioned in the introduction. However, because of using the analytic spectrum, the cepstral window does not have to depend on the data to preserve the known autocorrelations. In fact, it is a fixed rectangular window and, as a consequence, the algorithm to find $C^+(n)$ is extremely simple. On the other hand, the same spectral model (12) and, except by a factor 2, the same distance (11) are obtained using

$$T(\omega) = \frac{d}{d\omega} \log E(\omega) \qquad (13)$$

or

$$T(\omega) = -\tau(\omega) \qquad (14)$$

instead of (9) since these two expressions are, respectively, the real and imaginary parts of (9).

In speech recognition, the spectral distance between two frames will be evaluated with (11), where $C_1^+(n)$ and $C_2^+(n)$ are the corresponding cepstral coefficients which verify (12). Therefore, the summation in (11) will only extend from index 1 to index M.

SPEECH RECOGNITION RESULTS

Some preliminary speech recognition tests have been carried out with isolated words in a speaker dependent way to test the proposed method. The speech recognition system uses a pattern matching approach like Itakura's one [11] but with different paramerization and distance measure. After estimating the first 9 autocorrelations of every speech frame, the corresponding cepstral coefficients of the analytic spectrum are calculated with (7). The final result of the parameterization step are the values $nC^+(n)$ which are directly used in the Euclidean distance (11) between frames.

The vocabulary used in the test is formed by the ten Catalan digits. Each of ten speakers (7 male and 3 female) uttered ten repetitions of every digit. Fig. 1 shows the variance of $nC^+(n)$ along with the variance of $C^+(n)$, both evaluated on that data base. It is apparent that it will be preferable to use the unweighted Euclidean distance on $nC^+(n)$ rather than on $C^+(n)$ since the variance of the latter has a much greater dynamic range. This is another reason to choose the spectral distance given by (10) and (2) or (11) [12].

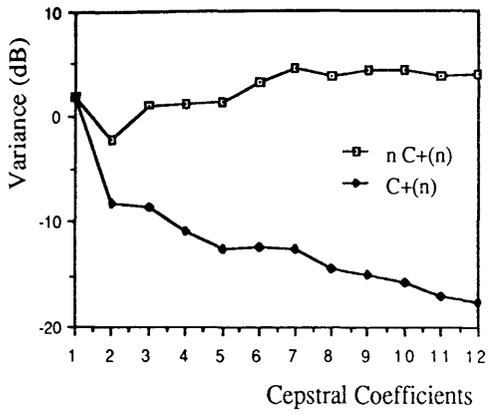

Fig. 1. Variances of $nC^+(n)$ and $C^+(n)$ in dB.

Recognition tests were performed, in a speaker dependent way, by taking one utterance per word as reference and testing the recognition system with the remaining nine repetitions. Then, the references were taken from another repetition and so forth. This means a total number of 9000 tests over the ten speakers. The LPC based system that uses the Itakura's distance [11] achieves an average recognition percentage of 99,5% that is greater than the 98,8% score obtained by the alternative spectral representation proposed here. However, the proposed technique is algorithmically and computationally simpler.

CONCLUSION

This paper is primarily devoted to present the analytic spectrum and its associated functions as possible tools in speech spectral representation and distance evaluation. These functions have the same poles as the spectrum and one of them, the spectral envelope $E(\omega)$, has already shown its usefulness in speech recognition when the signal is corrupted by noise [9].

In particular, we introduced a specific model for the analytic spectrum that, as the LPC all-pole model, arises from the Burg's optimization approach to spectral estimation. This model assumes that the analytic cepstrum is different from zero only in the indices where the autocorrelation is known. These non-zero values are obtained by means of a simple recursion.

In a coherent manner, the same Euclidean cepstral distance measure that is the object of the optmization is then used to calculate the spectral dissimilarity in speech recognition. Although the proposed technique is not yet fully developed, its recognition results are not far from those of the LPC technique and its computational load is lower.

REFERENCES

[1] L.R.Rabiner and S.E. Levinson, "Isolated and connected word recognition. Theory and selected applications", IEEE Trans. on Comm., Vol. COM-29, May 1981, pp. 621-59.

[2] J. Makhoul, "Linear prediction : a tutorial review", Proc of the IEEE, Vol. 63, No.4, Apr. 1975, pp. 561-580.

[3] J.P. Burg, Maximum Entropy Spectral Analysis, Ph.D. Dissertation, Stanford University, Stanford, CA 1975.

[4] C. Nadeu, "Maximum flatness spectral analysis", 4th. European Signal Process. Conf. EUSIPCO'88, Grenoble, Set. 1988.

[5] C. Nadeu, M. Bertran-Salvans and J. Solé, "Spectral estimation with rational modelling of the log spectrum", Signal Processing, 10 (1986), pp. 7-18.

[6] A.H. Gray and J.D. Markel, "Distance measures for speech processing", IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-24, pp. 380-391, Oct. 1976.

[7] M.A. Lagunas and M. Amengual, "Non-linear spectral estimation", Proc ICASSP'87, Dallas, Apr. 6-9, 1987, pp. 2035-8.

[8] A.V. Oppenheim and R.W. Schafer, Digital Signal Processing, Englewood Cliffs, NJ, Prentice-Hall, 1975.

[9] D. Mansour and B.H. Juang, "The short-time modified coherence representation and its application for noisy speech recognition", Proc. ICASSP'88, New York, Apr. 11-14, 1988, pp.525-528.

[10] D.H. Klatt, "Prediction of perceived phonetic distance from critical-band spectra: A first step",Proc.ICASSP'82,May1982, pp.1278-81.

[11] F. Itakura, "Minimum prediction residual principle applied to speech recognition", IEEE Trans. on Acoust. Speech and Signal Proces., Vol. ASSP-23, Feb. 1975, pp.67-72.

[12] Y. Tokhura, "A weighted cepstral distance measure for speech recognition",Proc. ICASSP'86, Tokyo,Apr.7-11,1987, pp.761-4.