# SPECTRAL REPRESENTATIONS FOR SPEECH RECOGNITION BY NEURAL NETWORKS — A TUTORIAL

B. H. Juang & L. R. Rabiner
AT&T Bell Laboratories
Murray Hill, New Jersey 07974

**Abstract. Recent advances in neural network classifier design have provided new insights into the problem of automatic speech recognition by machine. It has been shown that the speech representation used as the input to the network classifier is critically important for obtaining high recognition accuracy. In this paper we focus our attention on spectrum-based speech representations. Spectral representations, in order to be useful for speech recognition, need to be justified from both the computational (analytical) and the perceptual viewpoints. Our discussion of spectral representations, therefore, includes both the computational model and the associated measures of similarity that are appropriate for neural networks. This tutorial is intended to serve as a bridge between generic neural networks classifiers and classical speech analysis for speech recognition applications.**

## 1. INTRODUCTION

For speech recognition, a parsimonious representation of speech is essential and critical to overall system performance. Signal analysis methods, which convert speech into some type of parametric representation, are often the common denominator of all recognition systems, regardless of the particular classification approach taken in the design.

There exists a wide range of possibilities for representing the important properties of the speech signal. In this paper, we discuss representations that are based on the spectral properties of the speech signal, since spectral representations are arguably the most important parametric representations in speech recognition applications.

0-7803-0557-4/92$03.00 © 1992

Before giving a detailed discussion of various spectral representations, it is worth emphasizing the fact that a speech signal is a dynamic signal whose characteristics change with time in a particular manner and that linguistic isomorphism (e.g. two utterances that are considered as manifestations of the same word) does not imply acoustic isomorphism (e.g., the same word can be pronounced differently by different people). This acoustic variability is the origin of the fundamental difficulties that affect the derivation of "reasonable" speech representations. In this regard, a viable speech representation has to be built upon a linguistically as well as perceptually justifiable computational model. While this goal cannot totally be achieved, it should be understood in the following discussion that it is the ultimate objective of the research in this area.

## 2. SPECTRAL ANALYSIS MODELS

The two most common choices of spectral analysis model for speech recognition applications are a bank-of-filters model and an all-pole (linear prediction) model. The bank-of-filters model is shown in Fig. 1. The speech signal $s(n)$ is passed through a bank of $Q$ bandpass filters whose coverage spans the frequency range of interest in the signal. These filters generally overlap in

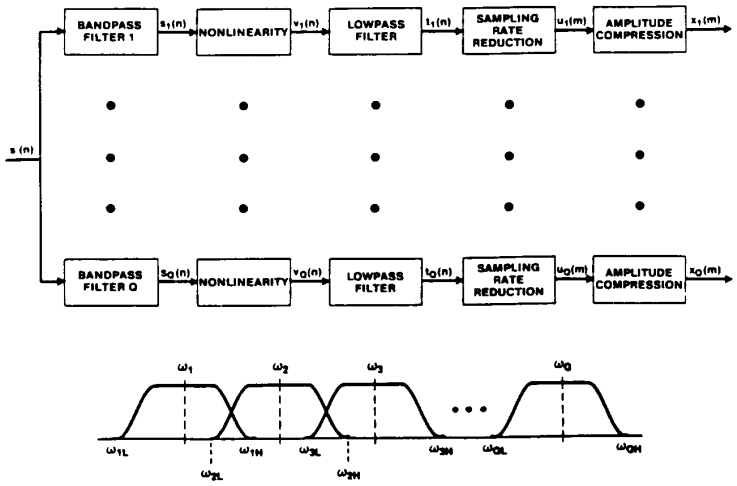


**Figure 1:**

A bank-of-filters spectral analysis model.

frequency as shown at the bottom of Fig. 1. The output of each channel is independently subjected to a specified nonlinearity as well as other signal processing to produce the spectral representation $X_n$. The nonlinearity is

typically a full wave or half wave rectifier, or a short time energy estimator. A low-pass filter is further used to obtain a smooth, slowly varying representation of the filtered output. This smooth spectral representation can then be resampled at a reduced rate (typically 50–100 times per second), producing a sequence of spectral vectors $\{X_n(m)\}$.

The LPC analysis approach performs spectral analysis on blocks of speech (speech frames) with an all-pole modeling constraint. This means that the spectral representation is constrained to be of the form $\sigma/A(e^{j\omega})$ where $A(e^{j\omega})$ is a $p$th order polynomial with $z$-transform

$$A(z) = 1 + a_1 z^{-1} + \ldots + a_p z^{-p} \; . \tag{1}$$

The order $p$ is called the LPC analysis order. The output of each LPC spectral analysis block is a vector of coefficients which define the spectrum of an all-pole model which best matches the signal spectrum over the period of time in which the frame of speech samples was accumulated.

## 2.1. Types of Filter Bank

The most common type of filter bank used for speech recognition is the uniform filter bank in which the center frequencies of the $Q$ bandpass filters are equally spaced to cover the frequency range of interest.

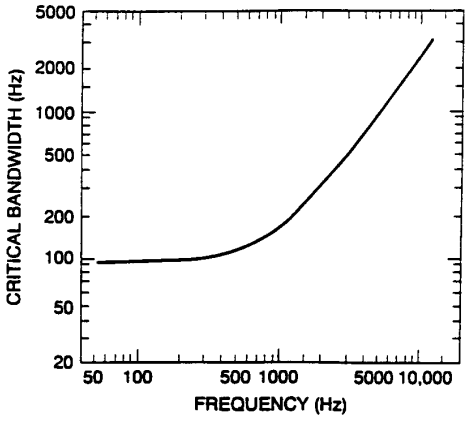


**Figure 2:**

The critical bandwidth as a function of the frequency at the center of the band.

An alternative to uniform filter banks is non-uniform filter banks designed according to some frequency spacing criterion. One commonly used criterion is to space the filters uniformly along a logarithmic frequency scale. Thus for

a set of $Q$ bandpass filters with center frequencies, $f_i$, and bandwidths, $b_i$, $i = 1, 2, \ldots, Q$, we set

$$b_1 = C \tag{2a}$$

$$b_i = \alpha b_{i-1}, \quad i = 2, 3, \ldots, Q \tag{2b}$$

and

$$f_i = f_1 + \sum_{j=1}^{i-1} b_j + \frac{(b_i - b_1)}{2} \tag{2c}$$

where $C$ and $f_1$ are the arbitrary bandwidth and center frequency of the first filter and $\alpha$ is the logarithmic growth factor. When $\alpha = 2$, the frequency spacing is an octave between bands.

An alternative to the octave band scale that is often used is the critical band scale [1]. The concept of a critical band is based on perceptual studies of speech articulation. It was shown that each critical band provided essentially equal contribution to speech intelligibility. Figure 2 shows the critical bandwidth as a function of the frequency at the center of the band. The use of a critical bandwidth filter bank leads to a scale for center frequency spacing which is approximately linear for frequencies below 1000 Hz and is close to logarithmic for frequencies above 1000 Hz. Similar perceptually motivated frequency spacing criteria lead to mel-scale and Bark-scale filter banks.

## 2.2. Linear Prediction Analysis

In all-pole modeling, a given speech sample $s(n)$ is assumed to be generated as a linear combination of the past $p$ samples and an excitation term, i.e.

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + Gu(n) \tag{3}$$

where $u(n)$ is a normalized excitation function, $G$ is the gain of the excitation and $a_1, a_2, \ldots, a_p$ are the prediction coefficients. A linear system model is shown in Fig. 3 where the transfer function $H(z)$ is an all-pole system of the form $1/A(z)$.

Given the speech signal, the predictor coefficients are calculated by minimizing the residual energy of the predicted speech signal. The residual signal is the difference between the actual speech signal and the predicted signal,

$$e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{k=1}^{p} a_k\, s(n-k) \tag{4}$$

and the residual energy is accordingly defined as

$$E = \sum_{n} e^2(n) \tag{5}$$

where the summation is over the (short-time) analysis interval (i.e. speech frame). The best set of predictor coefficients can be found by well known

methods such as the autocorrelation method or the covariance method [2]. The resultant spectral representation of the speech is thus $\sigma/A(e^{j\omega})$ which is defined by the coefficients $\{\sigma, a_1, a_2, \ldots, a_p\}$ where $\sigma \propto \sqrt{E_{\min}}$, where $E_{\min}$ is the minimum value of the prediction error energy.

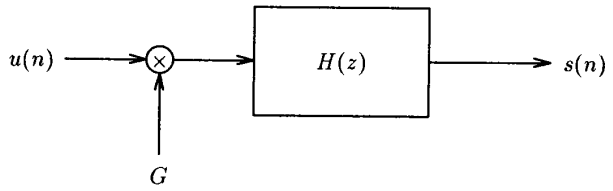


**Figure 3:**

A Linear System Model for Speech Prediction

## 3. SPECTRAL DISTORTION MEASURES AND PARAMETER TRANSFORMATION

A spectral representation, in order to be useful in speech recognition, needs to have associated with it a reasonable spectral distortion measure which gives a measure of the dissimilarity between two spectra. The concept of a good spectral representation, is, therefore, not an isolated one but is closely related to the way meaningful spectral dissimilarities are evaluated. Because of this relationship, there is a strong need to interpret a spectral representation in several transformed coefficient domains. This concept will become clear in this section.

### 3.1. Log Spectral Distances and Related Coefficient Transformations

Consider two power spectra $S(\omega)$ and $S'(\omega)$. The difference between the two spectra, on a log magnitude versus frequency scale is defined by

$$V(\omega) = \log S(\omega) - \log S'(\omega) \ . \tag{6}$$

One natural choice for a distance between $S$ and $S'$ is the set of $L_m$ norms defined by

$$d(S, S') = \left[\int_{-\pi}^{\pi} | V(\omega) |^m \frac{d\omega}{2\pi}\right]^{1/m} \ . \tag{7}$$

For $m = 2$, the distance is defined as the rms log spectral distortion which is widely used in many speech processing systems. It is well known that the

cepstrum, $c_n$, is the set of coefficients of the Fourier series representation of $\log S(\omega)$; i.e.

$$\log S(\omega) = \sum_{n=-\infty}^{\infty} c_n e^{-jn\omega} \ . \tag{8}$$

By applying Parseval's theorem, we can express the rms log spectral distortion in terms of the cepstrum:

$$d_2(S, S') = \left[ \sum_{n=-\infty}^{\infty} (c_n - c'_n)^2 \right]^{1/2} \tag{9}$$

where $c_n$ and $c'_n$ are the cepstra corresponding to $S$ and $S'$ respectively. Since evaluation of (9) is usually carried out with only a limited number of terms (normally$< 30$), the resultant distance computation is usually called the (truncated) cepstral distance.

The distance of (9) can be equally defined on cepstra derived from other spectral representations such as the filter bank output $\{X_n\}_{n=1}^{Q}$. When $X_n$ are obtained according to mel-frequency spacing, the resulting cepstrum is sometimes called a mel-frequency cepstrum or mel-cepstrum.

When $S(\omega)$ is modeled by a linear prediction model of the form $\sigma^2 / \mid A(e^{j\omega}) \mid^2$, the cepstral coefficients can be recursively computed:

$$\begin{aligned} c_0 &= \log \sigma^2 \\ c_n &= -a_n - \frac{1}{n} \sum_{k=1}^{n-1} k c_k a_{n-k} \quad \text{for} \quad n > 0 \end{aligned} \tag{10}$$

where $a_0 = 1$ and $a_k = 0$ for $k > p$, the linear prediction order. The cepstrum so derived is often called the "LPC cepstrum" to denote its difference from other cepstral representations.

For speech recognition, particularly speaker-independent speech recognition, further modification on the cepstrum is often desirable. It is well understood that the phonetic distinction among different speech sounds is most critically affected by the resonance structure of the spectrum but not the spectral tilt which is a strong function of the speaker's glottal characteristics among other factors. Also, higher cepstral coefficients have been shown to be highly susceptible to artifacts inherent in the spectral analysis method. Hence for a more reliable representation which contains fewer spurious components that are non-essential in speech recognition, liftering is often applied to the cepstral coefficients. Liftering is a weighting on the cepstrum; i.e. each $c_n$ is multiplied by a weight function $w(n)$ which can take various forms such as

$$w(n) = \begin{cases} 1 + h \sin\left(\dfrac{n\pi}{L}\right) & n = 1, 2, \ldots, L \\ 0 & n \leq 0, \quad n > L \end{cases} \tag{11}$$

where $h$ is normally $L/2$ and $L$ is typically $10 \sim 16$. Distance computations which use a liftered cepstrum are called "weighted cepstral distances".

## 3.2. Likelihood-Based Distortions

The linear prediction analysis can be formulated as a statistical estimation problem in which one seeks to maximize the likelihood of the prediction parameters for a given speech data frame. The difference in log likelihood can be considered a form of dissimilarity measure. The Itakura-Saito distortion, the Itakura distortion and the likelihood ratio distortion [3] are well known likelihood-based distortion measures.

A key component in the evaluation of likelihood-based distortions is the residual energy which can be expressed as

$$\begin{aligned} E &= \int_{-\pi}^{\pi} S(\omega) \mid A(e^{j\omega}) \mid^2 \frac{d\omega}{2\pi} \\ &= r(0)r_a(0) + 2\sum_{n=1}^{p} r(n)r_a(n) \end{aligned} \tag{12}$$

where $r(n)$ is the autocorrelation corresponding to $S(\omega)$ and

$$r_a(n) \triangleq \sum_{i=0}^{p-n} a_i a_{i+n} \quad \text{for} \quad n = 0, 1, 2, \ldots, p\,. \tag{13}$$

If the signal level, $\sigma$, is to be treated in a manner different from the spectral shape $A(e^{j\omega})$, normalization of the autocorrelation terms is necessary. For example, a residual normalized autocorrelation is often used in the evaluation of the likelihood ratio distortion between two *unity* gain all-pole spectra [3].

# 4. DYNAMIC SPECTRAL REPRESENTATIONS

The dynamic (temporal) behavior of the short-time spectrum of speech plays an important role in human perception of speech. Dynamic representations of speech are often derived by time differentiating the log spectrum. A first order differential spectrum is

$$\frac{\partial}{\partial t} \log S(\omega, t) = \sum_{n=-\infty}^{\infty} \frac{\partial c_n(t)}{\partial t} e^{-jn\omega} \tag{14}$$

where the temporal index $t$ of the short time spectral representations $S(\omega)$ and $c_n$ have been made explicit. The time derivative is normally obtained by polynomial approximation. Consider fitting a segment of the cepstral trajectory, $c_n(t)$, $t = -M, -M+1, \ldots, M$ by a second order polynomial $h_1 + h_2 t + h_3 t^2$. The fitting error $\sum_{t=-M}^{M}[c_n(t) - (h_1 + h_2 t + h_3 t^2)]^2$ can be minimized by choosing

$$h_2 = \sum_{t=-M}^{M} t\, c_n(t) \tag{15}$$

$$h_3 = \frac{T_M \sum_{t=-M}^{M} c_n(t) - (2M+1) \sum_{t=-M}^{M} t^2 c_n(t)}{T_M^2 - (2M+1) \sum_{t=-M}^{M} t^4} \tag{16}$$

and

$$h_1 = \frac{1}{2M+1} \left[ \sum_{t=-M}^{M} c_n(t) - h_3 T_M \right] \tag{17}$$

where $T_M = \sum_{t=-M}^{M} t^2$. The time derivatives of $c_n(t)$ can then be approximated by $h_2$ for the 1st order case and $2h_3$ for the 2nd order case. The resulting time derivatives are usually called the delta cepstrum and the delta-delta cepstrum respectively.

The dynamic representations can be considered supplementary features and have been demonstrated to be very effective in improving speech recognition performance.

## 5. NEURAL NETWORKS APPLICATIONS

In neural networks applications, particularly with hybrid learning schemes for pattern recognition [4], it is often desirable to categorize the inputs into clusters before supervised learning is performed to train the connection weights towards the output layer. Effectiveness of such clustering is a strong function of the speech representation and the associated distortion measure.

The various forms of the cepstral representation obviously are readily applicable in radial basis functions. In this case, the activation function for the $i$th node is a normalized Gaussian expression

$$g_i(\mathbf{c}) = \frac{\exp\left[-\|\mathbf{c} - \boldsymbol{\mu}_i\|^2 / 2\Sigma_i^2\right]}{\sum_k \exp\left[-\|\mathbf{c} - \boldsymbol{\mu}_k\|^2 / 2\Sigma_k^2\right]} \tag{18}$$

where $\boldsymbol{\mu}_i$ can be visualized as the centroid cepstral vector for node (cluster) $i$. (For simplicity, we have assumed that the elements of $\mathbf{c}$ are uncorrelated.) Therefore an input cepstral vector $\mathbf{c}$ close to $\boldsymbol{\mu}_i$ would produce a large response at node $i$. The "variance" $\Sigma_i^2$ indicates the degree of dispersion for the region associated with the $i$th cluster.

In a similar manner, the concept of radial basis functions can be applied when using likelihood related distortions (instead of the Euclidean distance in the exponents of (18)). A straightforward extension is to use

$$g_i(\mathbf{r}) = \frac{\exp(-E_i)}{\sum_k \exp(-E_k)} \tag{19}$$

where $\mathbf{r}$ is the input autocorrelation vector (dimension $p+1$) and the $E_k$'s are defined in the same way as Eq. (12), with $E_k$ representing the residual energy produced by the inverse filter $A_k(z)$ associated with cluster $k$.

## 6. SUMMARY

In this paper we have discussed various spectral representations suitable for speech recognition applications. These representations are intimately linked with appropriate spectral distortion measures that can be evaluated in the relevant domain of representation. We have also pointed out how these representations and spectral distortion measures can be applied in neural network solutions to pattern recognition problems.

## References


[1] E. Zwicker, G. Flottorp and S. S. Stevens, "Critical bandwidth in loudness summation," *J. Acoust. Soc. Am.* 29, pp. 548–557, 1957.

[2] J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*, Springer-Verlag, New York, 1976.

[3] R. M. Gray, A. Buzo, A. M. Gray, Jr. and Y. Matsuyama, "Distortion measures for speech processing," *IEEE Trans. ASSP-28*, no. 4, pp. 367–376, Aug. 1980.

[4] J. Hertz, A. Krogh and R. G. Palmer, *Introduction to the Theory of Neural Computation*, Addison-Wesley, Redwood City, CA, 1991.